Alternatives to Animal Testing: Research, Trends, Validation, Regulatory Acceptance

Jane Huggins
Toxicology Consulting Services, USA-Plainsboro

Content

1 Refinement of acute toxicity assay
1.1 Background
1.2 Refinement assays
1.3 Comparison of refinement assays
1.4 Humane endpoints
1.5 Regulatory activities
1.6 Summary, conclusions, and future work
2 Alternatives to eye corrosion/irritation testing in animals
2.1 Background
2.2 *In vitro* alternatives
2.2.1 Combinatorial approaches
2.2.2 Reference standards
2.2.3 Mechanistic considerations
2.3 Summary, conclusions, and future work
3 Alternatives to skin corrosion/irritation testing in animals
3.1 Background
3.2 *In vitro* models
3.3 Validation and regulatory activities
3.4 Summary, conclusions, and future work
4 Alternatives to skin sensitization testing in animals
4.1 Background
4.2 Alternative assays
4.3 Validation and regulatory activities
4.4 Summary, conclusions, and future work
5 Alternatives to developmental/reproductive toxicity testing in animals
5.1 Background
5.2 *In vitro* alternatives
5.2.1 Micromass (MM) assay
5.2.2 Whole embryo culture (WEC) assay
5.2.3 Embryonic stem cell (EST) test
5.2.4 Frog embryo teratogenesis assay (*Xenopus*) (FETAX)
5.3 Validation and regulatory activities
5.4 Summary, conclusions, and future work
6 Genetic engineering methodologies
6.1 Background
6.2 Genetically, engineered cell lines
6.3 Transgenic animals
6.4 Summary, conclusions, and future work
7 Gene chip technology as an alternative to animal testing
7.1 Background
7.2 What is a gene chip?
7.3 Application of gene chips to basic research
7.4 Application of gene chips to toxicological research
7.5 Development of gene chips as alternatives to animal testing
7.6 Current status of gene chip technology
7.7 Summary, conclusions, and future work
8 Validation of alternative methodology
8.1 Background
8.2 What is validation and why has it failed?
8.3 Improving the validation efforts
8.4 Harmonization of validation process
8.5 Why validate?
8.6 Mechanistic understanding: what are we validating?
8.7 Summary, conclusions, and future work

Summary

Current trends and issues in the development of alternatives to the use of animals in biomedical experimentation are discussed in this position paper. Eight topics are considered and include refinement of acute toxicity assays; eye corrosion/irritation alternatives; skin corrosion/irritation alternatives; contact sensitization alternatives; developmental/reproductive testing alternatives; genetic engineering (transgenic) assays; toxicogenomics; and validation of alternative methods. The discussion of refinement of acute toxicity assays is focused primarily on developments with regard to reduction of the number of animals used in the LD_{50} assay. However, the substitution of humane endpoints such as clinical signs of toxicity for lethality in these assays is also evaluated. Alternative assays for eye corrosion/irritation as well as those for skin corrosion/irritation are described with particular attention paid to the outcomes, both successful and unsuccessful, of several validation efforts. Alternative assays for contact sensitization and developmental/reproductive toxicity are presented as examples of methods designed for the examination of interactions between toxins and somewhat more complex physiological systems. Moreover, genetic engineering and toxicogenomics are discussed with an eye toward the future of biological experimentation in general. The implications of gene manipulation for research animals, specifically, are also examined. Finally, validation methods are investigated as to their effectiveness, or lack thereof, and suggestions for their standardization and improvement, as well as implementation are reviewed.

1 Refinement of acute toxicity assay

Three assays have been validated and adopted as replacements for the conventional LD_{50} test. The assays differ primarily as to

the endpoint they measure; however, all assays use fewer animals than the conventional LD_{50} test. The use of more humane endpoints, such as clinical signs of toxicity, rather than lethality, is perhaps the most advanced suggestion to date regarding toxicological evaluation of acute exposure. Much remains to be done, however, with regard to standardization of this approach.

2 Alternatives to eye corrosion/irritation testing in animals
Although much research has been done to date on the development of viable in vitro assays of ocular corrosion and irritancy, validation of these assays has been problematic. Several reasons have been postulated for the failure of validation efforts, the most prominent of which are the following: The in vivo test used for comparison, the Draize test, is based on subjective scoring of tissue lesions in the eye, providing variable estimates of eye irritancy; the non-animal method protocols were inadequate; the choice of test substances was not well-planned; and the statistical approaches used were not appropriate. Perhaps the most promising suggestion as to how to remedy these difficulties currently is to use complementary alternative assays in batteries to evaluate eye corrosion/irritation.

3 Alternatives to skin corrosion/irritation testing in animals
The use of alternative assays has replaced, to a large extent, the testing of corrosive or irritating substances on the skin of live animals. Examples of the assays discussed include the Corrositex® assay that uses no animal cells at all, the transcutaneous electrical resistance (TER) assay that uses a small section of rat skin, and several in vitro skin irritancy models that incorporate human skin in small quantities. From a scientific perspective, the replacement of the Draize test with these assays lends greater objectivity as well as more general relevance to human skin corrosion and irritation. Validation efforts utilizing these models have proven satisfactory in most instances.

4 Alternatives to skin sensitization testing in animals
Progress toward the development and refinement of alternative assays of contact sensitization is strongly dependent upon breakthroughs in our understanding of the immune processes mediating the response. Extensive efforts directed toward validation of the local lymph node assay have borne the much-needed fruit of a "stand alone" assay that incorporates elements of both refinement and reduction. However, much more basic research remains to be done before a fully validated replacement assay of contact sensitization finds regulatory support. Promising areas of research include those in which cytokine profiles associated with contact sensitization are analyzed.

5 Alternatives to developmental/reproductive toxicity testing in animals
Validation efforts are progressing well for in vitro assays of developmental/reproductive toxicity. Results from evaluations of the MM and WEC assays as well as the EST appear to be favorable; data from studies of FETAX suggest that further improvements in the assay would yield greater predictivity. Hence, our reticence to use an alternative assay to measure toxic effects on complex physiological processes such as reproduction may have to yield to the results obtained from these recent evaluations.

6 Genetic engineering methodologies
The generation and use of transgenic animals to study questions of biomedical interest have been questioned by many in view of the moral and ethical dilemmas presented by these activities. That transgenic animals may contribute to the reduction of animal use in toxicological experiments, particularly studies of carcinogenicity, is not disputed. However, advocates of replacement alternatives argue that in vitro alternatives to this type of toxicity testing have not been given adequate attention.

7 Gene chip technology as an alternative to animal testing
Gene chips (DNA microarrays) represent a technology that has already opened many doors in basic genomic research. Moreover, their value to both investigative and discovery toxicology is becoming much more apparent as more toxicology experiments are conducted using them. Use of microarrays as reduction or replacement alternatives to animal testing also holds great promise, particularly when they are used as components of prescreening batteries, and when coupled to cell culture techniques.

8 Validation of alternative methodology
Validation of alternative methods has just emerged from a rather chaotic phase in which the principles behind appropriate conduct of a validation study were defined, mainly through trial and error. Much refinement has come out of this "exploratory" phase, including recognition that validation studies should be built upon a solid platform, consisting of components such as good reference standards, reliable protocol transfer between laboratories, and appropriate application of biostatistical techniques. Efforts are now underway to apply these lessons learned to future validation studies and to harmonize validation techniques among countries in order to maximize the possibility that the data generated can be used worldwide.

Zusammenfassungen: Alternativen zu Tierversuchen: Forschung, Trends, Validierung, Akzeptanz auf behördlicher Ebene
In diesem Positionspapier werden aktuelle Trends und Ergebnisse in der Entwicklung von Alternativmethoden zu Tierversuchen in der biomedizinischen Forschung erörtert. Es werden acht Themenfelder angesprochen: Refinement akuter Toxizitätstests; Alternativen zu Augenkorrosion-/reizung; Alternativen zu Hautkorrosion-/reizung; Alternativen zu Kontaktempfindlichkeit gegenüber Stoffen; Alternativen zu Entwicklungs/-Reproduktionstests; Gentechnologie (transgen) Tests; Toxicogenomics und Validierung alternativer Methoden. Die Diskussion des Refinements akuter Toxizitätsuntersuchungen

konzentriert sich in erster Linie auf Entwicklungen im Hinblick auf eine Reduzierung der Tierzahlen beim LD_{50} Test. In diesen Untersuchungen wird ausserdem die Verwendung von menschlichen Endpunkten wie klinische Toxizitätsmerkmale für Letalität untersucht. Bei den Alternativmethoden im Bereich Augenkorrosionen/-reizungen sowie Hautkorrosionen/-reizungen wird das Hauptaugenmerk auf die Ergebnisse verschiedener Validierungsversuche gerichtet sein. Alternative Untersuchungen im Bereich Kontaktempfindlichkeit und Entwicklungs-/Reproduktionstoxikologie werden als Beispiele für Methoden dargestellt, welche zur Überprüfung von Interaktionen zwischen Toxinen und physiologisch komplexeren Systemen konzipiert wurden. Ausserdem werden die zukunftsträchtigen Bereiche Gentechnologie und Toxicogenomics besprochen. Die Auswirkungen von Genmanipulationen insbesondere auf Versuchstiere sollen untersucht werden. Abschliessend sollen Validierungsmethoden auf ihre Effektivität hin untersucht werden sowie Empfehlungen für deren Standardisierung, Verbesserung sowie Umsetzung überprüft werden.

1 Refinement von akuten Toxizitätstests
Bisher wurden drei Methoden für den Ersatz des LD_{50} Tests akzeptiert. Diese unterscheiden sich in erster Linie im von ihnen gemessenen Endpunkt; trotz dieses Unterschieds können durch diese Tests im Vergleich zum herkömmlichen LD_{50} Test Tiere eingespart werden. Heutzutage muss der vermehrte Einsatz von humanen Endpunkten, wie klinische Anzeichen von Toxizität, im Gegensatz zum Endpunkt Letalität als der fortgeschrittenste Ansatz hinsichtlich der toxikologischen Beurteilung akuter Exposition angesehen werden. Grosse Anstrengungen müssen jedoch in Hinblick auf die Standardisierung dieses Ansatzes unternommen werden.

2 Alternativen zu Tierversuchen im Bereich Augenkorrosion/-reizung
Obwohl bei der Entwicklung von in vitro Methoden im Bereich der Augenkorrosion/-reizung grosse Anstrengungen unternommen wurden, hat sich die Validierung der entsprechenden Methoden als problematisch herausgestellt. Verschiedene Gründe haben zu diesem Misserfolg beigetragen. Die bekanntesten Gründe sind die folgenden: Der für den Vergleich mit dem in vitro Test herangezogene in vivo Test, der Draize Test, beruht auf einer subjektiven Einteilung der auftretenden Gewebsverletzungen am Auge, was zu unterschiedlichen Einschätzungen der Augenreizung führt; das Protokoll der in vitro Methode war mangelhaft; die Auswahl der Testsubstanzen war schlecht geplant; die statistischen Analysemethoden waren ungeeignet. Als Lösung dieser Probleme bietet sich an, mittels in Batterie geschalteter verschiedener komplementärer Alternativmethoden Augenkorrosions/-reizungs Untersuchungen durchzuführen.

3 Alternativen zu Tierversuchen im Bereich Hautkorrosion/-reizung
Der Einsatz von Alternativmethoden hat in hohem Masse Tests auf korrosive oder reizende Substanzen der Haut an lebenden Tieren ersetzt. Im vorliegenden Dokument werden die folgenden Tests beschrieben: Corrositex®, bei welchem überhaupt keine Tierzellen eingesetzt werden; der transcutaneous electrical resistance (TER) Test, welcher einen kleinen Teil an Rattenhaut benötigt sowie verschiedene in vitro Hautreizungsmodelle, welche kleine Mengen menschlicher Haut erfordern. Aus wissenschaftlicher Sicht verhilft der Ersatz des Draize Tests durch die erwähnten in vitro Methoden im Bereich Hautkorrosion/-reizung zu mehr Objektivität sowie allgemeinerer Relevanz. Die mit diesen Modellen unternommenen Validierungsanstrengungen haben sich in vielen Fällen als zufriedenstellend erwiesen.

4 Alternativen zu Tierversuchen im Bereich Kontaktempfindlichkeit
Fortschritte in der Entwicklung und dem Refinement von Alternativmethoden im Bereich Kontaktempfindlichkeit hängen stark von den Erfolgen bezüglich unseres Verständnisses der Immunprozesse ab, welche die Reaktion vermitteln. Aus den grossen Anstrengungen, welche hinsichtlich der Validierung des local lymph node assay unternommen wurden, ist nun ein „stand alone" Test hervorgegangen, welcher sowohl zum Refinement wie zur Reduzierung von Tierversuchen beiträgt. Nichts desto trotz muss noch vermehrt Grundlagenforschung betrieben werden, bevor ein vollständig validierter Ersatztest für Kontaktempfindlichkeit Akzeptanz auf behördlicher Ebene erlangen wird. Erfolgversprechende Forschungsbereiche schliessen Analysen von Zytokinprofilen, welche mit Kontaktempfindlichkeit assoziiert sind, ein.

5 Alternativen zu Tierversuchen im Bereich Entwicklungs-/Reproduktionstests
Die Validierungsanstrengungen für in vitro Tests im Bereich Entwicklungs-/Reproduktionstoxizität kommen gut voran. Resultate der Evaluierung von MM und WEC Test sowie dem EST Test scheinen Erfolg versprechend zu sein; Daten von FETAX-Studien zeigen, dass eine weitere Verbesserung des Tests bessere Voraussagen ergeben würde. Daher sollte unsere Zurückhaltung, Alternativmethoden zur Messung toxischer Effekte auf komplexe physiologische Prozesse wie Reproduktion einzusetzen, zu den Resultaten führen, welche sich aus den früheren Evaluierungen ergeben haben.

6 Genmanipulation
Die Erzeugung für und der Einsatz von transgenen Tieren in der biomedizinischen Forschung hat bereits viel zu reden gegeben gerade hinsichtlich moralischer und ethischer Dilemmas, die sich aus solchen Aktivitäten ergeben. Dass transgene Tiere zu einer Reduzierung der Tierzahl in toxikologischen Experimenten, im Besonderen in der Krebsforschung beitragen können wird nicht bestritten. Dennoch sind die Verfechter von Ersatzmethoden der Meinung, dass den in vitro Methoden in diesem Bereich der Toxikologieprüfung zu wenig Beachtung geschenkt wird.

7 Gen chip Technologie als Alternative zu Tierversuchen
Gen chips (DNA Microarrays) repräsentieren eine Technologie, welche in der Genomforschung bereits viele Türen geöffnet hat.

Zudem wird ihre Bedeutung für die Toxikologie umso ersichtlicher, je mehr toxikologische Experimente unter deren Einsatz durchgeführt werden. Der Einsatz von Microarrays als Reduction oder Replacement Alternativen zu Tierversuchen muss als vielversprechend angesehen werden, insbesondere wenn diese als Komponenten von Präscreening Batterien eingesetzt und mit Zellkulturtechniken gekoppelt werden.

8 Validierung von Alternativmethoden
Die Validierung von Alternativmethoden entstand aus einer Zeit heraus, als die Richtlinien, welche einer angemessenen Durchführung einer Validierungsstudie zugrunde lagen, hauptsächlich über Versuch und Irrtum definiert wurden. Für das Refinement hat diese „Versuchsphase" viel gebracht, einschliesslich der Einsicht, dass Validierungsstudien einer soliden Grundlage entspringen sollten, welche aus Elementen wie gute Referenzstandards, zuverlässiger Protokolltransfer zwischen den Laboratorien und angemessener Einsatz biostatistischer Methoden bestehen. Es sind Anstrengungen im Gange, diese Erkenntnisse in zukünftige Validierungsstudien einfliessen zu lassen und Validierungstechniken zwischen den verschiedenen Ländern zu harmonisieren, um einen weltweiten Einsatz der generierten Daten zu ermöglichen.

Keywords: alternatives to animal use, 3R, validation, toxicity assays, eye corrosion, skin corrosion, contact sensitization, developmental/reproductive assays, genetic engineering, transgenics, toxicogenomics

1 Refinement of acute toxicity assays

1.1 Background

Historically, lethality following acute exposure to a chemical has been a cornerstone upon which much toxicological decision-making has rested. The LD_{50} (dose at which lethality is observed in 50% of the animals tested) is often considered the primary index of potential toxicity of a chemical and is widely used as a tool for determining the dosages to be used for further experimentation. LD_{50} values are derived using multiple species and routes of exposure. The most common species utilized is the rat, however, the mouse, guinea pig, rabbit, and dog are also tested. The most common routes of exposure are oral (by gavage), dermal, inhalation, and intraperitoneal and intravenous injection. Ideally, males and females of equal number per dose level are employed and several dose levels are evaluated. The LD_{50} value is obtained through probit analysis of the data obtained.

The LD_{50} test has come under attack for both ethical and scientific reasons because it uses a large number of animals, measures lethality as its major endpoint and produces variable results. Methods using fewer animals have been suggested as alternatives to the LD_{50} test (OECD, 1992; 1996; 1998). Moreover, incorporation of humane endpoints into animal testing has been advocated for reduction of animal pain and distress (OECD, 1999a).

1.2 Refinement assays

A continuum of refinement is noted in the three assays adopted as alternatives to the LD_{50} test. All three assays use fewer animals than the conventional LD_{50} test, all three assays emphasize humane treatment of animals undergoing testing, and one assay utilizes a major endpoint other than lethality as its determining value.

OECD (Organization for Economic Cooperation and Development) Guideline 423 describes the acute toxic class (ATC) method as follows: "This method is not intended to allow the calculation of a precise LD_{50}, but does allow for the determination of a range of exposures where lethality is expected since death of a proportion of the animals is still the major endpoint of this test. The results of the test should allow for classification according to any of the commonly used systems. Due to the sequential nature of the approach, the duration of the test could be longer than the procedure described in Test Guideline 401. The main advantage of this method is that it requires a smaller number of animals than both the "classical" acute oral toxicity (401) and the alternative fixed dose method (420). Moreover, because of the specific provisions for dose selection and interpretation, this method should increase consistency from laboratory to laboratory."

Both national and international validation studies have been conducted to evaluate the acute toxic class method as an alternative to the LD_{50} test (Schlede et al., 1992; Schlede et. al., 1995). Results from the national validation effort indicated that the method produced reliable results for the evaluation of toxicity and for classification of chemicals according to the classification system of the European Community (Tab. 1).

The ATC method also used substantially fewer animals than the LD_{50} test, and produced sufficient information about signs of toxicity. The participants in this study concluded that the ATC method could be applicable to routes of exposure other than oral, for example, dermal and inhalation. However, they felt that "because our present knowledge of signs of toxicity of substances with completely different chemical structures is limited and that obtaining 'sufficient reproducibility' of toxic signs is difficult, any approach not using death as the endpoint would be difficult to implement".

The international validation study of the ATC method utilized dosages consid-

Tab. 1: Results of the comparison of classification of the substances between the LD_{50} tests and the acute toxic class tests (from Schlede et al., 1992)

Nr.	Substance	Classification of LD_{50} data		Acute toxic class tests			
		based on the literature (contains reference numbers in brackets from original publication)	based on the estimated value	number of laboratories classifying a substance as			
				very toxic	toxic	harmful	unclassified
1	Aldicarb	very toxic (3)	very toxic	6	-	-	-
2	Parathion	very toxic (10)	very toxic	6	-	-	-
3	Di-isopropylfluorophosphate	very toxic (3)	very toxic	6	-	-	-
4	Thiosemicarbazide	very toxic (2)	very toxic	6	-	-	-
5	Indomethacin	very toxic (2)	very toxic	4	2	-	-
6	Phenylthiourea	very toxic (1)	very toxic	5	1	-	-
7	Mercury (II) oxide	very toxic (1); toxic (1)	toxic	1	5	-	-
8	Sodium arsenite	toxic (2)	toxic	1	5	-	-
9	Aldrin	toxic (6)	toxic	-	6	-	-
10	Allylalcohol	toxic (2)	toxic	-	6	-	-
11	Bis (tributyltin) oxide	toxic (4); harmful (1)	toxic	-	2	4	-
12	Acrylamide	toxic (3); harmful (1)	toxic	-	6	-	-
13	Cadmium chloride	toxic (1); harmful (2)	harmful	-	2	4	-
14	Methyl chloroformate	toxic (2); harmful (1)	harmful	-	2	4	-
15	Phenobarbital	toxic (1); harmful (2)	harmful	-	1	4	-
16	Caffeine	toxic (1); harmful (3)	harmful	-	1	5	-
17	Barium carbonate	harmful (3)	harmful	-	-	6	-
18	Aniline	harmful (4)	harmful	-	-	6	-
19	Ferrocene	harmful (3)	harmful	-	-	6	-
20	m-Dichlorobencene	harmful (1); unclassified (1)	harmful	-	-	4	2
21	Sodium salicylate	harmful (4)	harmful	-	-	6	-
22	Acetanilide	harmful (4)	harmful	-	-	6	-
23	Sodium lauryl sulphate	harmful (1); unclassified (1)	harmful	-	-	6	-
24	Acetonitrile	harmful (2); unclassified (3)	unclassified	-	-	2	4
25	Benzyl benzoate	harmful (1); unclassified (1)	unclassified	-	-	4	2
26	o-Phenylphenol	harmful (1); unclassified (2)	unclassified	-	-	1	5
27	Butylated hydroxyanisole	harmful (1); unclassified (6)	unclassified	-	-	-	6
28	N.N-Dimethyl formamide	unclassified (5)	unclassified	-	-	1	5
29	Quercetin dihydrate	unclassified (1)	unclassified	-	-	-	6
30	Ethylene glycol	unclassified (2)	unclassified	-	-	-	6

ered important for international harmonization of the method (i.e. 5, 50 and 500 mg/kg were used as well as 25, 200 and 2000 mg/kg). Nine laboratories from five countries participated in this study; twenty substances were tested. Findings from this study corroborated those noted in the national validation study. The lowest mean number of used animals was 6 or less and the highest mean number was 15. These numbers represent a reduction of 80% and 50%, respectively, in the number of used animals when compared to the LD_{50} test (30 animals). When the limit test with 2000 mg/kg is performed with the ATC method, six animals are used instead of ten animals that are used with the classical limit test. The number of moribund/dead animals per step in the ATC method was considerably less than that observed in a classical LD_{50} test (3 vs. 10-15). Hence, the ATC method was considered to subject fewer animals to pain and distress.

OECD Guideline 420 describes the fixed dose procedure (FDP) as follows: "Traditional methods for assessing acute oral toxicity, like Guideline 401, use death of animals as an endpoint. In 1984, a new approach to acute toxicity testing was suggested by the British Toxicology Society (BTS) based on a fixed dose procedure (British Toxicology Society, 1984). This avoided using death of animals as an endpoint, and relied instead on the observation of clear signs of toxicity developed at one of a series of fixed dose level. The fixed dose method set out in this guideline provides information both for hazard assessment purposes and for ranking substances. A preliminary sighting study, using a small number of

Tab. 2: A comparison of the classification allocated to the test substances by the LD_{50} and fixed dose tests (from Van den Heuvel et al., 1990).
*Four laboratories allocated "very toxic" on the basis of a dose ranging study only

			Fixed dose tests – number of laboratories classifying compound as:			
Compound		**LD_{50} test classification**	Very toxic	Toxic	Harmful	Unclassified
A	Nicotine	Toxic	-	23	3	-
B	Sodium pentachlorophenate	Harmful	-	1	25	-
C	Ferrocene	Harmful/unclassified	-	-	3	23
D	2-Chloroethyl alcohol	Toxic	-	19	7	-
E	Sodium arsenite	Toxic	-	25	1	-
F	Phenyl mercury acetate	Toxic	2	24	-	-
G	p-Dichlorobenzene	Unclassified	-	-	-	26
H	Fentin hydroxide	Toxic	-	8	17	1
J	Acetanilide	Harmful	-	-	4	22
K	Quercetin dihydrate	Unclassified	-	-	-	26
L	Tetrachlorvinphos	Unclassified	-	-	1	25
M	Piperidine	Harmful	-	2	24	-
N	Mercuric chloride	Toxic	-	25	1	-
P	1-Phenyl-2-thiourea	Toxic harmful	12	12	2	-
R	4-Aminophenol	Harmful	-	-	17	9
T	Naphthalene	Unclassified	-	-	-	26
U	Acetonitrile	Harmful	-	-	4	22
W	Aldicarb (10%)	Very toxic	22*	-	-	-
X	Resorcinol	Harmful	-	-	25	1
Y	Dimethyl formamide	Unclassified	-	-	-	26

animals, is included in this guideline in order to estimate the dose effect for toxicity and mortality and to provide information on dose selection for the main study. Results from the sighting and main studies enable compounds to be ranked in different classification systems, currently in use."

Van den Heuvel et al. (1990) conducted an international validation study on the fixed dose procedure as an alternative to the classical LD_{50} test. 33 laboratories in 11 countries evaluated the toxic effects of 20 substances using the fixed dose procedure and compared these effects to those obtained using the classical LD_{50} test (Tab. 2).

This investigation produced consistent results that were not substantially affected by inter-laboratory variations and provided adequate information on signs of toxicity including their nature, time to onset, duration and outcome. Fewer animals than the OECD guideline for acute toxicity testing (401) were used and animals were subjected to less pain and distress than the classical LD_{50} test. Utilization of this method enabled substances to be ranked according to the EEC classification system. However, this validation effort also highlighted the variability between laboratories established in different countries in assessments of signs of toxicity on which a decision to intervene and humanely kill animals is based. As noted in the validation studies of the ATC method, the investigators of the FDP concluded that the principle of the procedure was clearly applicable to acute toxicity testing by dermal or inhalation routes.

OECD Guideline 425 describes the up-and-down procedure (UDP) as follows: "This test procedure is of principal value in minimizing the number of animals required to estimate the acute oral toxicity of a chemical and in estimating a median lethal dose. The median lethal dose allows for comparison with historical data. In addition to the observation of mortality, it allows the observation of signs of toxicity. The latter is useful for classification purposes and in the planning of additional toxicity tests."

Bruce (1985) developed the UDP by first conducting a historical review of a large number of conventional acute toxicity studies. These studies were used as a basis for choosing length of time between successive doses, the sex of the animals to be tested and the spacing between doses in the UDP. The second investigation was a computer simulation based on data contained in the historical dataset. The results from the simulation were in excellent agreement with the historical data indicating that the UDP could be used as an alternative to the classical LD_{50} test. Subsequently, the Bruce procedure was adopted by the American Society for Testing and Materials (ASTM, 1987). Bruce suggested that the UDP offered substantial savings in numbers of animals although he indicated that estimated LD_{50} values will be less precise than those obtained from larger experiments. Moreover, he cautioned that the method may be inap-

propriate for chemicals typically producing animal death two or more days after administration.

Three validation studies of the UDP procedure have been conducted in which the ability of the UDP to estimate the LD_{50} was compared to that obtained using the traditional method described in Testing Guideline 401 (Bruce, 1987; Bonnyns et al. 1988; Yam et al., 1991). For all 25 chemicals evaluated, the average ratio of the LD_{50}'s for the two methods compared was 1.76. These data indicate that the two methods essentially provide the same point estimate of the LD_{50} for the chemicals tested (Fig. 1).

1.3 Comparison of refinement assays

When the up-and-down and fixed dose procedures were compared against the classical LD_{50} test by Yam et al. (1991), both methods were found to reduce the numbers of animals used while providing adequate information for ranking the 10 materials tested according to the European Economic Commission classifications for acute oral toxicity. The signs observed and the duration of signs tended to vary among methods. The authors concluded that while different doses used in the three methods probably accounted for most of the observed differences in signs, they also considered that laboratory variations in sign recording may have also contributed to the observed differences. In total, for the 10 test materials, the classical LD_{50} test generated 67 signs, the up-and-down method generated 62 signs, and the fixed dose procedure generated 49 signs. Comparing the fixed dose procedure and the up-and-down method with regard to autopsy findings resulted in the finding that the fixed dose procedure produced fewer autopsy findings. This was not surprising, since the fixed dose procedure generally used lower doses than the up-and-down method. Both alternatives used fewer animals than the classical method. By using females, the up-and-down method required only 50% as many animals as the fixed dose procedure, and 29% as many as the classical method. The fixed dose procedure produced the fewest deaths of the three tests.

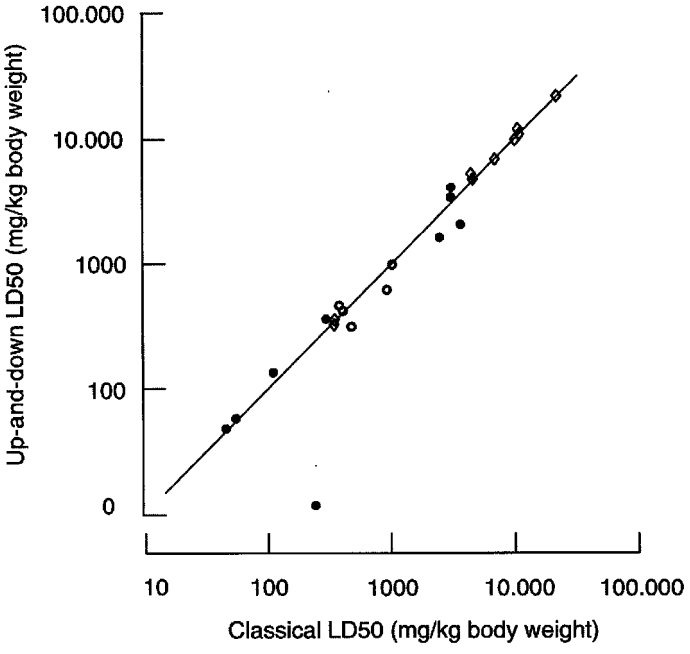

Fig. 1: Comparison of the LD_{50} determined using the up-and-down method with the LD_{50} estimated from conventional tests for materials testes by van den Heuvel (1990) and Yam et al. (1991) (●), Bruce (1987) (◊) and Bonnyns et al. (1990) (○) (from Lipnick et al., 1995).

Further comparisons of the up-and-down and fixed dose procedures, and conventional LD_{50} test were performed by Lipnick et al. (1995). The authors' major conclusions that favor the UDP are as follows:

- The UDP generally produces an estimate of the LD_{50} that is similar to that achieved from conventional acute toxicity testing.
- Data on chemicals tested in the UDP lead to the same EEC acute toxicity classification as do those from the conventional LD_{50} test in 23 of 25 reviewed cases. These results are as good as those for the FDP *vs.* conventional LD_{50} test where 16 out of 20 classifications are coincident. For seven out of 10 cases, the UDP and FDP lead to the same classification.
- The UDP gives an estimate of the LD_{50} and thus data from this test method are applicable to any acute toxicity classification system. In contrast, FDP data are directly referable to the classification system used by the EEC. However, by use of the information from the sighting study for the FDP, classification decisions can be made for other reasons.
- Testing with the UDP requires only between 6 and 10 animals of one sex, the smallest number of animals of any protocol. In contrast, the FDP usually uses 10 or 20 animals, while the conventional LD_{50} determination generally requires 30 animals (15 if only one sex is used). Moreover, the OECD protocols for the conventional test and FDP call for a sighting study which uses up to another five animals, a sighting study is not needed for the UDP.
- To date, the UDP has been used to evaluate lethality as an endpoint. Given that the frequencies of toxic manifestations are similar for the chemicals that have been simultaneously investigated in the UDP and the FDP (72% and 64%, respectively), it seems reasonable to explore further the applicability of the UDP to non-lethal toxicity endpoints.
- Analyses conducted here, and a review of the literature, indicate that the two sexes usually respond similarly in acute oral toxicity tests. When responses differ, females are generally

more sensitive than males. Consideration should be given to restricting acute toxicity testing of chemicals to females unless there is information suggesting that males are more sensitive for a given substance.

1.4 Humane endpoints

Russell and Burch (1959) defined refinement as any development leading to a "decrease in incidence or severity of inhumane procedures applied to those animals which have to be used". Hence, incorporation of humane endpoints into animal testing protocols adds considerable refinement to these studies. Animals undergoing testing for endpoints such as tumor production, infectious disease, vaccine potency, and target organ toxicity are treated more humanely by such a measure and studies in which lethality is the major endpoint measured can actually be replaced by those in which other signs of toxicity are monitored. However, as noted by the researchers below, clear definition of humane endpoints and development of the methods by which they can be assessed effectively are obviously necessary.

Morton (2000) describes a systematic approach for establishing humane endpoints. He advocates the use of score sheets that list the cardinal clinical signs that are observable and measurable, and the key clinical signs are identified through the experience of those involved in the research. He suggests that lists of clinical signs be developed by very closely observing the first few animals undergoing a new scientific procedure. The list is modified with experience until a set of signs is established that most animals will show during that experiment and that are relevant to the assessment of pain and distress. These cardinal signs are set out against time in the score sheet. Use of these score sheets encourages closer observation of animals by all involved at critical times in the experiment, subjective assessments are avoided to a large extent, and consistency of scoring is increased.

Toth (2000) has advanced a data-based approach for predicting imminent death and defining specific moribund conditions in objective terms. She indicates that the moribund state can be defined by identifying the values of various variables that precede imminent death that can serve as "signals" for preemptive euthanasia. She stresses that specific variables should be identified and weighted in terms of their predictive value. However, she acknowledges that objective data-based criteria that predict imminent death may not always fit comfortably into the goals of an experiment. Hypothermia, inability to rise or ambulate, weight loss, and biochemical variables are all suggested as potential predictors of imminent death.

Schlede et al. (2000) discuss specifically the use of humane endpoints in acute oral toxicity testing. Their evaluation of clinical signs was made in rats used for validation studies of the acute toxic class method. These data demonstrated that all forms of "convulsions" resulted in death in 94% (484/516) of rats, and the "lateral position" resulted in death in 79% (177/223) of rats. Clinical signs associated with a high mortality rate in this study are listed in Table 3.

Wallace (2000) discusses humane endpoints in cancer research and makes the following suggestions:

- Tumor growth or excision should replace "survival" endpoints.
- Many preparatory procedures (e.g. low-level whole body irradiation, immunosuppressive agents, surgical ablation of endocrine glands) may represent a greater challenge to an animal that than of a developing tumor; hence, humane endpoints should consider the cumulative effect of all experimental challenges.
- Tumor development and animal condition should be monitored frequently because unexpected or uncontrolled tumor development can result in unnecessary animal distress or mortality.

Olfert and Godson (2000) propose that increases in serum levels of cytokines be used as indicators of the presence of infectious disease and as predictors of both onset and outcome of infectious disease. This proposition is supported by the fact that changes in the levels of these parameters occur early in the disease process, before severe behavioral and physiologic changes do. Additionally, body weight change, weight loss, and decreased activity are reflective of changes in cytokine level. The authors suggest that changes such as these are all measurably more humane endpoints than is allowing progression of the infectious disease within the animal model.

Dennis (2000) writes that "death is not usually intended in genetic engineering studies, but lethality or animals with severe health problems are commonly encountered. Genetically-engineered

Tab. 3: Clinical signs associated with a high mortality rate (from Schlede, Gerner and Diener, 2000).

Clinical sign	Number of rats	Dead/moribund rats	%
Convulsion			
-Convulsion (unspecified)	43	43	100
-Clonic convulsion	218	207	95
-Tonic convulsion	96	79	82
-Tonic-Clonic convulsion	125	122	98
-Saltatory convulsion	10	10	100
Lateral position	223	177	79
Ventral position	9	9	100
Tremor	389	296	76
Gasping	143	108	76
Vocalisation	97	79	81
Extension spasm	6	6	100
Flexion spasm	8	8	100
Coma	9	9	100
Decrease of muscle tone	18	16	89
Mucoid faeces	35	27	77

animals often have a decreased ability to resist disease, increased tumor production, or compromised basic bodily functions such as eating or breathing". He suggests that it is extremely important that institutions supervise and continually review ongoing studies to identify problems as they occur and to ensure that appropriate humane endpoints are established.

These comments all indicate that humane endpoints can be incorporated into many diverse toxicological protocols. Perhaps the most encouraging regulatory support for doing so is found in guidelines promulgated by the OECD which summarize the use of clinical signs as humane endpoints for experimental animals used in safety evaluation. Included in these guidelines is a listing of types of effects that should be monitored in an adequate evaluation of an animal to determine its condition and whether there might be evidence indicative of pain and or distress:

- Changes in physical appearance (e.g. coat texture; hair soiled with urine or faeces)
- Changes in clinical signs (e.g. respiration rate; posture)
- Changes in unprovoked behavior (e.g. self mutilation; compulsive behavior)
- Behavioral changes in response to external stimuli (e.g. excitability; righting reflex)
- Changes in body weight, and related changes in food and water consumption
- Changes in clinical parameters (e.g. body temperature, heart and respiration rate, clinical chemistry and hematology).

This listing concurs with many of the suggestions given by the researchers above. Moreover, the guiding principles of this OECD document include the statement that, "severe pain, suffering, or death are to be avoided as endpoints". Hence, continued usage of any test where lethality is the endpoint appears to be in conflict with these guidelines.

1.5 Regulatory activities

The body of information about refinement alternatives to the conventional LD_{50} test has reached "critical mass" in that three alternative assays have been extensively studied and deemed appropriate as replacements for the conventional LD_{50} test by members of the scientific community and, to a certain extent, by those of the regulatory community as well. Furthermore, incorporation of humane endpoints (which are defined in at least one case as those which do not include death) into toxicity testing is being strongly encouraged by members of both communities. In December 2002, the Test Guideline 401 (the conventional LD_{50} test) has been deleted and replaced by alternative methods of acute toxicity testing (OECD, 1999b). In order to accomplish this objective, several revisions in the three alternative assays available need to be made. Despite the fact that the UDP is the only test that provides a point estimate of the LD_{50}, it does not provide estimates of the slope of the dose-response curve and confidence interval that are needed by regulatory agencies in some instances. Therefore, these variables need to be included in a revised procedure. Moreover, both the FDP and ATC method need to be changed to reflect changes in the regulatory classification scheme brought about by recent global harmonization efforts.

The USEPA has agreed to revise the UDP to include a procedure that would provide slope and corresponding confidence interval estimates. Accordingly, a revised UDP has been undergoing expert review by the Interagency Coordinating Committee on Validation of Alternative Methods (ICCVAM) (USEPA, 2000a). The revised procedure includes a modified up-and-down procedure that improves performance, a modified limit test that utilizes only females and provides a limit dose of 5000 mg/kg for specific regulatory purposes, and an added supplemental test for determining the slope and confidence interval. The panel's review of this revised procedure has been completed (July 25, 2000); their recommendations are to be posted by the end of January, 2001 (USEPA, 2000b).

1.6 Summary, conclusions, and future work

In December 2002 the Test Guideline 401 (conventional LD_{50} test) has been deleted. Three assays have been adopted as replacements for the conventional LD_{50} test. The assays differ primarily as to the endpoint they measure: the ATC method and FDP provide ranges of values that are applicable to particular regulatory classification schemes; whereas, the UDP provides an actual point estimate of the LD_{50} value. However, all of these assays use fewer animals than the conventional LD_{50} test.

In addition to assays using fewer animals, refinement of acute toxicity testing is being supported by such documents as the recent set of guidelines issued by the OECD for the use of clinical signs as humane endpoints in toxicity testing. These guidelines suggest strongly that lethality is no longer an acceptable endpoint, hence, they support substitution of an ED_{50} value for the LD_{50} value in acute toxicity testing. Furthermore, researchers reviewing the refinement assays proposed as alternatives to the conventional LD_{50} test suggest that humane endpoints can be incorporated effectively into those assays.

This recent, rapid progress toward full regulatory acceptance of alternative assays which refine acute toxicity testing represents an exciting chapter in the history of the alternatives to animal testing movement. However, in view of the fact that the ultimate goal of this research is replacement of animals in acute toxicity testing, further work will be focused toward validation of true replacement alternatives (i.e. those that do not use animals.) Exemplary research efforts in this area have been those by the Multicenter for the Evaluation of *In Vitro* Cytotoxicity (MEIC). The results of this work have just been published (Clemedson and Ekwall, 1999; Ekwall, 1999). They were the focal point of a public meeting sponsored by ICCVAM that investigated alternative assays for predicting acute systemic toxicity in order to lay a framework for further regulatory acceptance (ICCVAM, 2000).

Worldwide acceptance and incorporation of refinement assays such as, and including those discussed above should be seen in the next five years, if not sooner judging from recent events. Batteries of cytotoxicity tests should also be used much more frequently for prediction of the acute toxicity of new chemical com-

pounds during this time frame. The only significant hurdle remaining to be cleared is that of acceptance of an endpoint other than lethality. Our reliance on the LD_{50} as the endpoint of choice is, in many respects, simply a product of prior conditioning as well as a lack of feasible alternatives, neither of which should impede our progress now.

By the end of the next ten years, cytotoxicity assays should have been researched fully enough to provide data to support validation of one or more of these methods as replacements for the use of animals in acute toxicity testing. By the end of the next twenty years, regulatory acceptance of cytotoxicity or other similar endpoint as a reliable indicator of acute toxicity should be in evidence. Moreover, the mechanistic events linking cytotoxicity in cell culture to lethality in the whole organism should be well-defined.

Future research and funding efforts should be directed toward more precise methods of defining humane endpoints as well as validation of cytotoxicity assays as replacements for animal use in acute toxicity testing. Although much concerned thought and action has been directed toward the incorporation of humane endpoints into toxicity protocols, further work (as noted by all the investigators cited above) is definitely needed in order to bring definition and resolution to observations of clinical signs. Systematic methods of observation and effective in-depth training of animal-handling personnel are critical to the successful implementation of humane endpoints, particularly if the data are to be used in a quantitative fashion. Moreover, work in which patterns of toxic sign development are studied in relation to both the animal model used and the chemical(s) administered should be supported.

Many methods currently exist for the measurement of cytotoxicity endpoints. Efforts should be made to assess which are the most cost-effective, reliable indicators of acute toxicity when compared to either animal LD_{50} or human acute lethality data or, perhaps, to both. This type of validation will prove invaluable to the development of effective predictive batteries of these methods as well as to their eventual acceptance by regulatory bodies. The work performed by the MEIC is a good starting point for this type of endeavor because it provides comparative information among many different types of cytotoxicity assay.

References

ASTM (American Society for Testing and Materials) (1987). *Standard test method for estimating acute oral toxicity in rats*. Designation: E 1163-87. In Annual Book of ASTM Standards, Philadelphia.

Bonnyns, E., Delcour, M. and Vral, A. (1988). Up-and-Down Method as an Alternative to the EC-Method for Acute Toxicity Testing. IHE Project No. 2153/88/11. Institute of Hygiene and Epidemiology, Ministry of Public Health and the Environment, Brussels.

British Toxicology Society (1984). Special report: A new approach to the classification of substances and preparations on the basis of their acute toxicity. *Hum. Toxic. 3*, 85-92.

Bruce, R. (1985). An up-and-down procedure for acute toxicity testing. *Fund. Appl. Toxicol. 5*, 151-157.

Bruce, R. (1987). A confirmatory study for the up-and-down method for acute toxicity testing. *Fund. Appl. Toxicol. 8*, 97-100.

Clemedson, C. and Ekwall, B. (1999). Overview of the final MEIC results: I. The in vitro-in vivo evaluation. *Toxic. In Vitro 13*, 657-663.

Dennis, M. (2000). Humane endpoints for genetically engineered animal models. *ILAR Journal 41(2)*, 94-98.

Ekwall, B. (1999). Overview of the final MEIC results: II. The in vitro-in vivo evaluation, including the selection of a practical battery of cell tests for prediction of acute lethal blood concentrations in humans. *Toxic. In Vitro 13*, 665-673.

ICCVAM (Interagency Coordinating Committee for the Validation of Alternative Methods) (2000). International Workshop on In Vitro Methods for Assessing Acute Systemic Toxicity. October 17-20. Arlington, VA.

Lipnick, R., Cotruvo, J., Hill, R. et al. (1995). Comparison of the up-and-down, conventional LD50, and fixed-dose acute toxicity procedures. *Fd. Chem. Toxic. 33(3)*, 223-231.

Morton, D. B. (2000). A systematic approach for establishing humane endpoints. *ILAR Journal 41(2)*, 80-86.

OECD (Organization for Economic Cooperation and Development) (1987). Guideline No. 401 – Acute Oral Toxicity.

OECD (1992). Guideline No. 420 – Acute Oral Toxicity: Fixed Dose Procedure.

OECD (1996). Guideline No. 423 – Acute Oral Toxicity: Acute Toxic Class.

OECD (1998). Guideline No. 425 – Acute Oral Toxicity: Up and Down Procedure.

OECD (1999a). Guidance Document on Humane Endpoints for Experimental Animals Used in Safety Evaluation Studies. Paris: OECD.

OECD (1999b). OECD Document ENV/JM (99) 19, Test Guidelines Programme, Acute Oral Toxicity Testing: Data Needs and Animal Welfare Considerations, 29th Joint Meeting, June 8-11, Paris, France.

Offert, E. and Godson, D. (2000). Humane endpoints for infectious disease animal models. *ILAR Journal 41(2)*, 99-104.

Russell, W. and Burch, R. (1959). The Principles of Humane Experimental Technique. London: Methuen & Co. LTD. (Reissued: 1992, Universities Federation for Animal Welfare, Herts, England).

Schlede, E., Mischke, U., Roll, R. and Kayser, D. (1992). A national validation study of the acute-toxic-class method – An alternative to the LD_{50} test. *Arch. Toxicol. 66*, 455-470.

Schlede, E., Mischke, U., Diener, W. and Kayser, D. (1995). The international validation study of the acute toxic class method (oral). *Arch. Toxicol. 69*, 659-670.

Schlede, E., Gerner, I. and Diener, W. (2000). The use of humane endpoints in acute oral toxicity testing. In M. Balls, A.-M. Zeller and M. E. Halder (eds.), *Progress in the Reduction, Refinement and Replacement of Animal Experimentation (907-914)*. Amsterdam, London, New York, Tokyo: Elsevier 11.

Toth, L. (2000). Defining the moribund condition as an experimental endpoint for animal research. *ILAR Journal 41(2)*, 72-79.

USEPA (United States Environmental Protection Agency) (2000a). Notice (65 FR 08385): Request for Data and Nomination of Expert Panel of Scientists to Participate in the Independent Peer Review Evaluation of the Revised Up-and-Down Procedure for Assessing Acute Oral Toxicity. *Federal Register Volume 65(34)*. February 18.

USEPA. (2000b). Notice (65 FR 35109) of Peer Review Meeting on the Revised Up-and-Down Procedure (UDP) as an Alternative Test Method for Assessing Acute Oral Toxicity; Request for Comments. *Federal Register Volume 65 (106)*. June 1.

Van den Heuvel, M., Clar, D., Fielder, R. et al. (1990). The international validation of a fixed-dose procedure as an alternative to the classical LD_{50} test. *Fd. Chem. Toxic. 28(7)*, 469-482.

Wallace, J. (2000). Humane endpoints and cancer research. *ILAR Journal 41(2)*, 87-93.

Yam, J., Reer, P. and Bruce, R. (1991). Comparison of the up-and-down method and the fixed-dose procedure for acute oral toxicity testing. *Fd. Chem. Toxic. 25*, 259-263.

2 Alternatives to eye corrosion/irritation testing in animals

2.1 Background

Current testing guidelines for eye corrosion/irritation testing promulgated by OECD (Organization for Economic Cooperation and Development) and USEPA (United States Environmental Protection Agency) use the following definitions for corrosion and irritation. Eye corrosion is defined as the "production of irreversible tissue damage in the eye following application of a test substance to the anterior surface of the eye". Eye irritation is defined as the "production of reversible changes in the eye following the application of a test substance to the anterior surface of the eye". (OECD, 1987; USEPA, 1998). These test guidelines also incorporate the following humane considerations which support the use of alternatives to eye corrosion/irritation testing:

1. "Strongly acidic or alkaline substances, for example, with a demonstrated pH of 2 or less or 11.5 or greater, need not be tested owing to their predictable corrosive properties. Buffer capacity should also be taken into account.
2. Materials which have demonstrated definite corrosion or severe irritation in a dermal study need not be further tested for eye irritation. It may be presumed that such substances will produce similarly severe effects in the eyes.
3. Results from well validated and accepted *in vitro* test systems may serve to identify corrosives or irritants such that the test material need not be tested *in vivo*."

Furthermore, the number of animals used in testing is limited by these guidelines as follows: "A single animal should be considered if marked effects are anticipated. If the results of this test in one animal suggest the test substance to be a severe irritant (reversible effect) or corrosive (irreversible effect) to the eye using the procedure described, further tests may not need to be performed. In cases other than a single animal test, at least three animals should be used. Occasionally, further testing in additional animals may be appropriate to clarify equivocal responses."

Moreover, a draft revised version of OECD Guideline 405 includes the recommendation that an integrated testing strategy for a stepwise evaluation of all existing information on the substance including, e.g., data from human experience and from *in vitro* tests be incorporated (OECD, 2000). Hence, much regulatory support of alternatives to eye corrosion/irritation testing in animals is in evidence. Furthermore, many alternative models for eye corrosion/irritation testing have been developed. Unfortunately, validation of these models has remained elusive. A recent report by ECVAM (European Center for the Validation of Alternative Methods) discusses several potential reasons for this lack of validation and puts forth suggested initiatives for remedy (Balls et al., 1999). These include the use of reference standards (RS), stepwise testing strategies, multivariate and other statistical techniques for the further analysis of data generated in previous validation studies, and a program of mechanistic research.

2.2 *In vitro* alternatives

Historically, alternatives to eye corrosion/irritation testing in animals were considered the most important assays to develop in view of the unquestionable pain and suffering experienced by animals upon which severely irritating and corrosive substances were tested. Numerous *in vitro* assays have been developed over the last twenty years in response to this need. Some of the more well-known tests include the bovine corneal opacity and permeability (BCOP) assay (Gautheron et al., 1992), the hen's egg test – chorioallantoic membrane (HET-CAM) assay (Lopke, 1986), and several cytotoxicity tests (for example, 3T3-neutral red uptake (3T3-NRU)) (Borenfreund and Borrero, 1984; Borenfreund and Puerner, 1985).

2.2.1 Combinatorial approaches

As mentioned above, a number of efforts directed toward validation of a single *in vitro* assay of eye corrosion/irritation have been conducted without considerable success. However, when multivariate analysis was applied to the results from these studies, it indicated that assays used together in complementary fashion may provide good predictive information. Multivariate analysis was

applied to results obtained from the European Commission/British Home Office (EC/HO) validation study (Balls et al., 1995). The analysis revealed that combinations of data from assays of epithelial integrity (fluorescein leakage (FL) test), *ex vivo* models (isolated rabbit eye (IRE), isolated chicken eye (ICE)) and a cytotoxicity test (neutral red uptake (NRU)) explained more of the variability in the data than any single test used alone. This finding resulted in calculation of a better prediction model (PM). Similarly, when multivariate analysis of data obtained from a validation study conducted under the auspices of COLIPA (European Cosmetic, Toiletry and Perfumery Association) (Bagley et al., 1997) was performed, results comparable to those obtained in the EC/HO study were obtained (i.e. improved PMs could be developed based on combinations of *in vitro* endpoints). Furthermore, a validation study was coordinated by the Centre for Documentation and Evaluation of Alternative Methods to Animal Experiments (ZEBET) at the Bundesinstitut für Risikobewertung (BfR), and supported financially by the German Department of Research and Technology (BMBF) (Spielmann et al., 1993; Spielmann et al., 1996). Results from multivariate analysis of this study indicated that chemicals could be reliably classified as severe irritants (classification R41) through the combined use of the HET-CAM test and the 3T3 NRU test.

Hence, analysis (and re-analysis) of results from several validation studies suggests that complementary pairing of *in vitro* assays (usually a cytotoxicity test with an organotypic test) can be considered useful to the prediction of eye corrosion/irritation. Independent research investigations of the predictivity of *in vitro* assays used in combination (batteries) have also yielded good results (for example, Lewis et al., 1994; Pham and Huff, 1999; Rosenkranz and Cunningham; 2000).

The phrase "combinatorial approach" can be applied not only to combining complementary *in vitro* assays into batteries of tests, but also to combining data from many different types of experiments, not just *in vitro* assays. When a large combination of different data types is analyzed, this procedure is often referred to as "tier-testing", "hierarchical testing" or as a "stepwise strategy". Results from the complementary pairing of *in vitro* assays can contribute much to hierarchical testing schemes such as the stepwise strategy currently suggested in the OECD revised 405 guidelines. The following points from these revised guidelines outline the types of data to be considered prior to *in vivo* testing, including results from *ex vivo* and *in vitro* assays.

- Existing human or animal data: When there is sufficient human data from the test substance, it may not need to be tested in animals.
- Structure activity relationships (SAR). Historical experience (including human data) or testing of structurally related chemicals should be evaluated. If there are sufficient data to indicate the eye irritancy/corrosivity potential of a chemical or mixture from analogues, the test substance can be presumed to produce similar responses. SAR experiences should be interpreted cautiously when evaluating non-irritating/ corrosive substances.
- Physicochemical properties and chemical reactivity. Strongly acidic or alkaline substances which can be expected to result in a pH in the eye of 2 or less, or 11.5 or greater, may not need to be tested owing to their probable corrosive properties. Buffering capacity (alkaline or acidic reserve) should also be taken into consideration.
- Results from skin irritation studies. Substances that have demonstrated severe skin irritancy or corrosivity in a single application dermal study may not need to be tested for eye irritancy and corrosion. It can be presumed that such substances will produce similar severe effects on the eyes.
- Results from other studies. If a substance is highly toxic by the dermal route, it need not be tested in the eyes because it can be assumed to be highly toxic by this route as well.
- Results from *in vitro* or *ex vivo* tests that are generally accepted for purposes of hazard or risk assessment. Substances that have demonstrated the potential in an *in vitro* or *ex vivo* study to be corrosive or a severe irritant may not need to be tested for irritation and corrosion *in vivo*. It can be presumed that such substances will produce similar severe effects on the eyes.
- If there is insufficient evidence with which to evaluate the potential eye irritation/corrosivity of a substance from the preceding information, a skin irritation/corrosion test (see Guideline 404 and its Attachment) should be performed first. If the substance is shown to produce severe skin irritation or corrosion, it can be presumed that it would also produce similar effects in the eyes, so that an *in vivo* eye test need not be performed.

ECVAM has evaluated the stepwise strategy suggested by OECD in its revised guidelines for acute eye irritation/corrosion testing and concluded that the strategy is effective in reducing and refining the use of the Draize eye test (Worth and Fentem, 1999).

2.2.2 Reference standards

Balls et al. (1999) have suggested that a reference standards approach be used in validation studies of *in vitro* assays of eye corrosion/irritation. They emphasize that, "the term 'reference standard' (RS) should not be confused with 'positive control'". Rather, they define a positive control as a substance which is known to give a positive response and which is used to confirm the correct conduct of the assay. Alternatively, a reference standard is a substance which has a known degree of toxicity *in vivo*, and which can be used *in vitro* to determine the degree of toxicity of test substances, whose effects are scaled relative to the RS. This group also hypothesizes that the reference standard approach to eye corrosion/irritation testing *in vitro* will include the following roles:

- within companies, for the development and cross-validation of *in vitro* assays
- in the validation of alternative methods, as a replacement for the totally blind approach which currently exists, so that substances can be grouped into categories defined by the reference standards
- in regulatory toxicology, for the submission of data on selected new substances to authorities

An evaluation of the use of reference standards in the validation process has begun by the ECVAM Reference Standards Working Group. Five *in vitro* tests of eye corrosion/irritation have been nominated as candidates for analysis. These include the ICE, BCOP, HET-CAM, NRU (neutral red uptake), Epi-Ocular™, and RBC (red blood cell) hemolysis assays. Conduct of this evaluation will involve testing of reference standards from different chemical groups, development of a PM based on the results from the reference standards, and application of the derived PM to a second set of chemicals, the identities of which are unknown. This work, if successful, should lay a much needed foundation for reliable evaluation of results from validation efforts in terms of comparisons to chemicals of known toxicity.

2.2.3 Mechanistic considerations

A recent article by Bruner et al. (1998) highlights the importance of understanding the mechanisms of eye irritation, "particularly when attempting to improve *in vitro* prediction of *in vivo* eye irritancy". Efforts by ECVAM to evaluate the failure of validation studies of eye corrosion/irritation have also pinpointed understanding mechanisms of action behind eye corrosion/irritation as critical to any future validation/acceptance of *in vitro* assays of this insult (Balls et al., 1999).

Consideration of underlying mechanisms of action is in evidence in much of the research and development of *in vitro* assays of eye corrosion/irritation. For example, recognition of the importance of mechanisms of action is implicit in the use of complementary assays. Single assays of cytotoxicity or organotypic effects probably do not explain the entire mechanism of action behind development of corrosion or irritation. Furthermore, measurement of different endpoints within the same assay has proven to be valuable in discriminating mechanisms of action. An excellent example of this is the BCOP assay in which Gautheron et al. (1992) realized the importance of measuring more than one indicator of irritation in the bovine cornea. Hence, the BCOP assay investigates both opacity and permeability.

A recent workshop held in Brussels, Belgium in October, 1998 as a follow-up to the work discussed by Bruner above suggested the following areas as foci for mechanistic research of eye corrosion/irritation:

- development of an appropriate set of reference test substances for use in the research
- evaluation of the area and depth of corneal injury as markers of eye injury
- exploration of the use of early biomarkers of eye injury (for example, cytokine release)
- development of methods for evaluating corneal wound healing
- development of methods for assessing the kinetics of eye injury
- development of methods for assessing injury to nerve cells in the cornea

Development of an appropriate set of reference test substances is currently underway at ECVAM by the Working Group on Reference Standards. However, much work remains to be done on the remaining five areas of suggested mechanistic research indicated above.

Evaluation of the area and depth of corneal injury has been addressed by several researchers. Although this work was performed *in vivo*, it is thought to have important ramifications for the development of *in vitro* alternatives for ocular irritancy. Jester et al. (1996) investigated the application of *in vivo* confocal microscopy (CM) to the understanding of surfactant-induced ocular irritation. The aim of this research was to "assess the ability of *in vivo* confocal microscopy to provide noninvasively derived histopathologic correlates of surfactant-induced eye irritation from which specific pathologic mechanisms could be identified". Rats and rabbits, received anionic or cationic surfactant in one eye with the other eye serving as control. Eyes were examined and scored for ocular irritancy subsequently using a penlight and slit-lamp. Corneas were then evaluated by *in vivo* CM to evaluate epithelial layer thickness and surface epithelial cell area, corneal thickness, depth of necrosis, inflammation, fibrosis, and endothelial injury. The anionic surfactant produced slight irritation (peak scores of 12.4 and 8.0) and *in vivo* CM revealed changes limited to the corneal epithelium.

Maurer and colleagues (1997) probed the uses of CM microscopy further. Surfactants of slight, mild, moderate, and severe irritancy were directly applied to the corneas of rabbits and eyes and eyelids were examined macroscopically and scored for irritation beginning 3 hour after dosing and periodically through day 35. Concurrently, the corneas were evaluated by *in vivo* CM. Three-dimensional data sets extending from the surface epithelium to the endothelium were assessed for surface epithelial cell size, epithelial layer thickness, total corneal thickness and depth of keratinocyte necrosis. Results indicated that significant differences in area and depth of injury occur with surfactants of differing irritancy. Furthermore, the data suggested that differences at 3 hours can be used to distinguish different levels of ocular irritation. Application of CM microscopy to evaluation of the irritancy of unknown surfactants has also been performed (Maurer et al., 1998). Macroscopic and microscopic findings regarding the ocular irritation of six surfactants of relatively unknown irritancy were compared to those of six surfactants of known irritancy. The right eye of each rat tested received the surfactant directly on the cornea. Untreated left eyes served as controls. At 3 hours and on days 1, 3, and 35, eyes and eyelids were collected for microscopic examination. Macroscopic and microscopic findings indicated that three surfactants were similar to mildly irritating surfactants and three were similar to moderately irritating surfactants previously studied.

The premise that cytokines can be used as an early biomarker of corneal injury is supported by Planck (1999) who explained their role in this way: "Infection and tissue damage activate nearby cells to produce a group of proteins, called cytokines, which help mediate the resulting inflammatory and repair processes. Interleukin-1 (IL-1), IL-6, and tumor necrosis factor alpha (TNF1) are often called master cytokines because they are produced by many cell types and have multiple effects on target cells including synthesis of additional cytokines. All

three cellular layers of cornea appear capable of producing and responding to these master cytokines, although the regulation of production and the repertoire of responses are not clear. Knowledge of the roles of these master cytokines in response to corneal insults should enhance the development of methods to manipulate repair and inflammatory processes therapeutically."

Cytokine release has been investigated as a measure of irritancy in dermal cells, considered by some researchers to represent various aspects of ocular tissue. Co-cultures of human dermal fibroblasts with human epidermal keratinocytes and human dermal fibroblasts in three-dimensional culture have both been used as *in vitro* assays of ocular irritation in which inflammatory response was measured with cytokines (Curren et al., 1997).

Future research will undoubtedly include definitive studies identifying the cytokines involved in ocular response to injury and the kinetics of their action. This work should be greatly aided by the use of functional human corneal equivalents constructed from cell lines (Griffith et al., 1999). These equivalents comprised the three main layers of the cornea (epithelium, stroma, and endothelium). Each cellular layer was fabricated from immortalized human corneal cells that were screened for morphological, biochemical, and electrophysiological similarity to their natural counterparts. Equivalents mimicked human corneas in key physical and physiologic functions, including morphology, biochemical marker expression, transparency, ion and fluid transport, and gene expression. Work performed by Sotozono et al. (1997) profiles cytokine expression following injury to mouse corneas. Although performed *in vivo*, this evaluation and others like it, could be used on a comparative basis with research performed *in vitro*, perhaps using human corneal equivalents. Development of methods for evaluating corneal wound healing is benefiting from evaluations of ocular repair mechanisms performed both *in vivo* and *in vitro*. Although some of the experiments were not conducted with the express purpose of developing an alternative to animal testing, the methodologies used support such a goal. For example, Burgalassi and co-workers (2000) evaluated the effect of xyloglucan (tamarind seed polysaccharide (TSP)) on conjunctival cell adhesion to laminin and on corneal epithelium wound healing. Cultured human conjunctival cells were labeled by addition of a tritiated amino acid mixture. Their adhesion to laminin-coated culture wells in the absence or presence of TSP was checked by radioactivity count. TSP was also tested *in vivo* in animals with corneal damage. Compared to hyaluronate, TSP slightly but significantly increased the wound healing rate *in vivo*. TSP (1%) also exerted a positive influence on cell adhesion to laminin, up to a certain laminin concentration. These researchers concluded that the ability of the polysaccharide to promote corneal wound healing might depend on its influence on the integrin-substrate recognition system.

In another study of corneal healing, Lambiase et al. (2000) studied the effects of nerve growth factor (NGF) on corneal repair in human and rat corneal epithelial cells in culture and human corneal organ culture. They showed that NGF is a constitutive molecule present and produced in normal human and rat corneas and that *in vitro* human and rat corneal epithelial cells, produce, store, and release NGF, and also express high-affinity NGF receptors. In human organ culture, epithelium, keratinocytes, and endothelium were shown to bind exogenous radiolabeled NGF, and epithelial cell binding was increased after epithelium

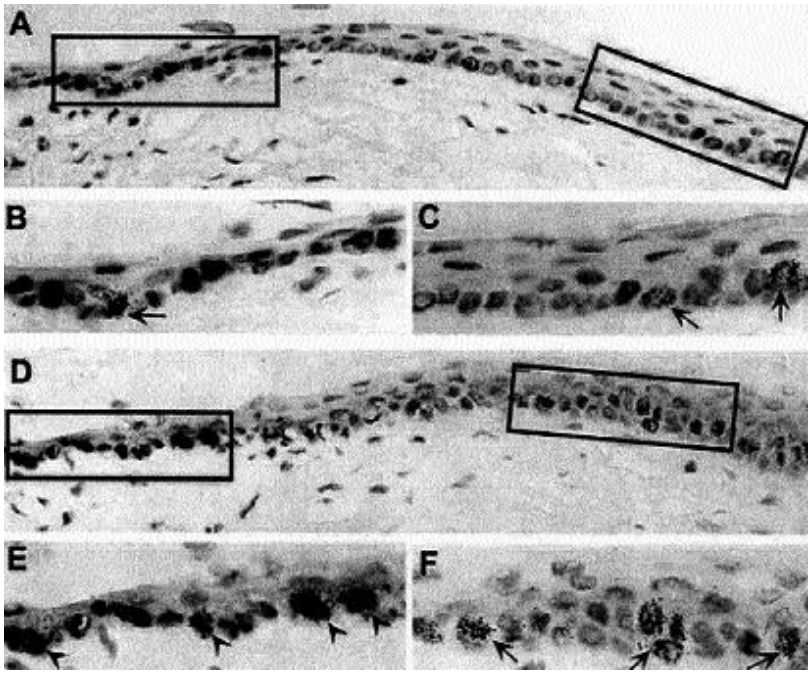

Fig. 2: Stem cells located in limbal epithelium can be rapidly induced to enter the proliferative population.
Long term labeling with BrdU to detect slow cycling stem cells (LRCs; red stained nuclei) followed by a single pulse of ^{3}H-TdR to detect rapidly cycling TA cells (arrows) demonstrates that under resting conditions (A, B) all slow-cycling cells are preferentially located in the limbus, while most TA cells are located in the peripheral corneal epithelium (A, C). An occasional TA cell can also be observed among the limbal epithelial stem cells (arrow, B). Twenty-four hours following n-heptanol-induced central corneal wound (D, E, F), a single pulse of ^{3}H-TdR was administered to mice that had populations of LRCs (red stained nuclei). Many of the LRCs were now double-labeled (arrowheads, E) indicating that they had incorporated ^{3}H-TdR and thus were undergoing a round of DNA synthesis. In addition, there was an increase in TA cells in the peripheral corneal epithelium (F, arrows) suggestive that this population also was induced to proliferate in response to wounding (from Lehrer et al., 1998).

injury. The authors concluded that, "NGF plays an important role in corneal physiopathology".

In a study of the kinetics of corneal wound healing, Lehrer and co-workers (1998) studied the replication of corneal epithelial stem cells, and their progeny, transit amplifying (TA) cells. Using double labeling techniques, they showed that the stem cells can be induced to enter DNA synthesis by wounding. They found that TA cells of the peripheral cornea undergo at least two rounds of DNA synthesis whereas those of the central cornea are capable of only one round of division in response to wounding (Fig. 2). Moreover, cell cycle time of transit amplifying cells can be shortened and number of times these cells replicate is increased in response to wounding. These results could contribute to evaluation of endpoints derived from *in vitro* models, such as the corneal equivalent discussed above.

Development of *in vitro* assays that will evaluate damage to the corneal nerve may benefit from recent work in assay development in the more general field of *in vitro* neurotoxicity. For example, the acute neurotoxic effects of trimethyltin (TMT) have been quantified using neuronal networks cultured on microelectrode arrays (Gramowski et al., 2000). Spontaneously active monolayer networks *in vitro* were cultured on thin film microelectrode arrays. Two different types of mouse CNS tissues exhibited "characteristic and dose-dependent changes of their electrophysiological activity patterns after treatment with TMT". Moreover, rat cortical neuron cultures have been used to differentiate the activities of structurally diverse chemicals (e.g. 2,5-hexanedione, acrylamide, organophosphates) (Schmuck et al., 2000). Effects on cytoskeletal elements and on the energy state of the cells were used as endpoints as well as cytotoxicity.

Neurological endpoints have been measured previously using corneal tissue cultured *in vitro.* Mikulec and co-workers (1995) used *in vitro* rabbit cornea preparations for both electrophysiological recording and wound healing measurements after treatment with the corneal analgesic, diltiazem, a calcium channel blocker. An earlier study by Tanelian and MacIver (1990) used rabbit corneal tissue isolated and maintained *in vitro* to facilitate staining, visualization, and electrophysiologic recording of corneal nerves (Fig. 3).

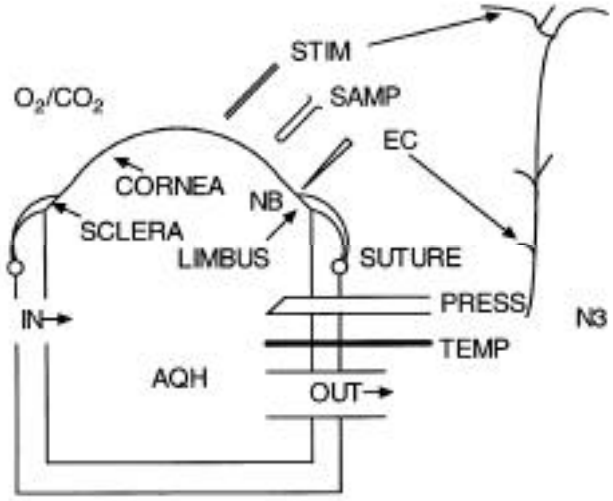

Fig. 3: Diagram of the isolated cornea preparation showing placement of stimulating probe (STIM) on free nerve endings (FNE) terminals, micropipette used to apply dye samples (SAMP) and extracellular recording electrode (EC) placed on a nerve bundle (NB).
The sclera was sutured to a plexiglas ring forming a pressure tight seal. Artificial aqueous humor solution (AQH) was continuously perfused (1.0 ml/min) and temperature (TEMP; 35°C) ad pressure (PRESS; 18mm Hg) monitored (from Tanelian and MacIver, 1990).

This work investigated the effects produced by two methylpyridinium fluorescent dyes on electrophysiologic responses from corneal A-delta and C fiber afferents. Nerve fibers were selectively stained by the dyes and could be followed from their point of entry in small nerve bundles at the cornea-sclera border to individual free nerve ending terminals in the corneal epithelium. Hence, investigative methods, both old and new, can perhaps be employed in the search for a better mechanistic understanding of corneal nerve injury.

2.3 Summary, conclusions, and future work

Current regulatory initiatives favor and emphasize the need to incorporate alternatives to eye corrosion/irritation testing in animals into the testing of new products for safety. Although much research has been done to date on the development of viable *in vitro* assays of ocular corrosion and irritancy, acceptable validation of these assays has not been achieved. Several reasons have been postulated for the failure of validation efforts, the most prominent of which are the following. The *in vivo* test used for comparison, the Draize test, is based on subjective scoring of tissue lesions in the eye, providing variable estimates of eye irritancy; the non-animal method protocols were inadequate; the choice of test substances was not well-planned; and the statistical approaches used were not appropriate (Balls et al., 1999).

Several suggestions have been made as to how to remedy these difficulties. Because no single *in vitro* assay of ocular corrosion/irritation can be said to encompass fully the *in vivo* response, use of complementary assays in batteries may prove to be a more suitable approach. Re-evaluation of results from several validation efforts seems to support this suggestion as do results from independent studies. Application of more discriminating methods of statistical analyses have also contributed to a much better understanding of the relationships

between/among various *in vitro* assays when used in combination. Use of complementary *in vitro* assays is also supported by the current emphasis on tier-testing strategies that incorporate more than one level/type of data in decision-making structures.

Finally, many voices (past and present) have recommended that we define the mechanisms behind the production of eye corrosion/irritation. Studies of the reference standards used, extent of corneal damage, early biomarkers, corneal wound healing, kinetics of corneal insult and repair and corneal nerve injury have all been suggested as major areas of research to be pursued vigorously in the next decade.

For the present, prompt action in delineating which assays of ocular corrosion/irritation to use in complementary fashion will aid the swift incorporation of those assays into the testing of new products through tier-testing schemes. Re-evaluation of past validation efforts coupled with newer initiatives aimed at definition of a reference standards approach should yield firmer ground upon which to plan future validation studies. These efforts will be needed, particularly as assays of mechanisms of action become more prevalent and in need of validation.

Alternatives to eye corrosion/irritation testing in animals are the oldest alternatives in existence and they are perhaps the assays from which we have learned the most. Current re-evaluation efforts as well as those focused on development of newer, mechanistically-oriented models can only serve to contribute more to our understanding of eye corrosion/irritation and to the ultimate replacement of animals in this type of toxicity testing.

References

Bagley, D. M. et al. (1997). A summary report of the COLIPA international validation study on alternatives to the Draize rabbit eye irritation test. *Toxicol. In Vitro 11*, 141-179.

Balls, M., Botham, P. A., Bruner, L. H. and Spielmann, H. (1995). The EC/HO international validation study on alternatives to the Draize eye irritation test. *Toxicol. In Vitro 9*, 871-929.

Balls, M., Berg, N., Bruner, L. et al. (1999). Eye irritation testing: The way forward. The report and recommendations of ECVAM workshop 34. *ATLA 27*, 53-77.

Borenfreund, E. and Borrero, O. (1984). In vitro cytotoxicity assays: potential alternatives to the Draize ocular irritancy test. *Cell Biol. Toxicol. 1*, 55-65.

Borenfreund, E. and Puerner, J. A. (1985). Toxicity determined in vitro by morphological alterations and neutral red absorption. *Toxicol. Lett. 24*, 119-124.

Bruner, L. H., de Silva, O., Earl, L. K. et al. (1998). Report on the Colipa workshop on mechanisms of eye irritation. *ATLA 26*, 811-820.

Burgalassi, S., Raimondi, L., Pirisino, R. et al. (2000). Effect of xyloglucan (tamarind seed polysaccharide) on conjunctival cell adhesion to laminin and on corneal epithelium wound healing. *Eur. J. Ophthalmol. 10(1)*, 71-76.

Curren, R. D., Sina, J. F., Feder, P. et al. (1997). IRAG Working Group 5: Other assays. Interagency Regulatory Alternatives Group. *Food Chem. Toxicol. 35(1)*, 127-158.

Gautheron, P., Dukic, M., Alix, D. and Sina, J. F. (1992). Bovine corneal opacity and permeability test: an in vitro assay of ocular irritancy. *Fundam. Appl. Toxicol. 18(3)*, 442-449.

Gramowski, A., Schiffmann, D. and Gross, G. W. (2000). Quantification of acute neurotoxic effects of trimethyltin using neuronal networks cultured on microelectrode arrays. *Neurotoxicology 21(3)*, 331-342.

Griffith, M., Osborne, R., Munger, R. et al. (1999). Functional human corneal equivalents constructed from cell lines. *Science 286 (5447)*, 2051-2053.

Jester, J. V., Maurer, J. K., Petroll, W. M. et al. (1996). Application of in vivo confocal microscopy to the understanding of surfactant-induced ocular irritation. *Toxicol. Pathol. 24*, 412-428.

Lambiase, A., Manni, L., Bonini, S. et al. (2000). Nerve growth factor promotes corneal healing: structural, biochemical, and molecular analyses of rat and human corneas. *Invest. Ophthalmol. Vis. Sci. 41(5)*, 1063-1069.

Lehrer, M. S., Sun, T. T. and Lavker, R. M. (1998). Strategies of epithelial repair: modulation of stem cell and transit amplifying cell proliferation. *J. Cell Sci. 111(Pt 19)*, 2867-2875.

Lewis, R. W., McCall, J. C. and Botham, P. A. (1994). Use of an in vitro test battery as a prescreen in the assessment of ocular irritancy. *Toxicol. In Vitro 8*, 75-81.

Lopke, N. P. (1986). HET (Hen's Egg Test) in toxicological research. In R. Marks and G. Plewig (eds.), *Skin Models. Models to Study Function and Disease of Skin* (282-291). Wien, Heidelberg, New York: Springer-Verlag.

Maurer, J. K., Li, H. F., Petroll, W. M. et al. (1997). Confocal microscopic characterization of initial corneal changes of surfactant-induced eye irritation in the rabbit. *Toxicol. Appl. Pharmacol. 143*, 291-300.

Maurer, J. K., Parker, R. D. and Carr, G. J. (1998). Ocular irritation: pathological changes occurring in the rat with surfactants of unknown irritancy. *Toxicol. Pathol. 26*, 226-233.

Mikulec, A. A., Lukatch, H. S., Monroe, F. A. and MacIver, M. B. (1995). Diltiazem spares corneal A delta mechano and C fiber cold receptors and preserves epithelial wound healing. *Cornea 14(5)*, 490-496.

OECD (Organization for Economic Cooperation and Development) (1987). Guideline for Testing of Chemicals No. 405: Acute Eye Irritation/Corrosion.

OECD (2000). Draft Revised Guideline for Testing of Chemicals No. 405: Acute Eye Irritation/Corrosion.

Pham, X. T. and Huff, J. W. (1999). Cytotoxicity evaluation of multipurpose contact lens solutions using an in vitro test battery. *Clao J. 25*, 28-35.

Planck, S. R. (1999). Cytokines in corneal injury responses. Crisp Data Base National Institutes of Health (CRISP/99/EY11921-01A1).

Rosenkranz, H. S. and Cunningham, A. R. (2000). A battery of cell toxicity assays as predictors of eye irritation: A feasibility study. *ATLA 28(4)*, 603-607.

Schmuck, G., Ahr, H. J. and Schluter, G. (2000). Rat cortical neuron cultures: an in vitro model for differentiating

mechanisms of chemically induced neurotoxicity. *In Vitro Mol. Toxicol. 13(1)*, 37-50.

Sozotono, C., He, J., Matsumoto, Y. et al. (1997). Cytokine expression in the alkali-burned cornea. *Curr. Eye Res. 16(7)*, 670-676.

Spielmann, H., Kalweit, S., Liebsch, M. et al. (1993). Validation study of alternatives to the Draize eye irritation test in Germany: cytotoxicity testing and HET-CAM test with 136 industrial chemicals. *Toxicol. In Vitro 7*, 505-510.

Spielmann, H., Liebsch, M., Kalweit, S. et al. (1996). Results of a validation study in Germany on two in vitro alternatives to the Draize eye irritation test, the HET-CAM test and the 3T3 NRU cytotoxicity test. *ATLA 24*, 741-858.

Tanelian, D. L. and MacIver, M. B. (1990). Simultaneous visualization and electrophysiology of corneal A-delta and C fiber afferents. *J. Neurosci. Methods 32(3)*, 213-222.

USEPA (United States Environmental Protection Agency) (1998). Health Effects Test Guidelines. OPPTS 870.2400. Acute Eye Irritation. August.

Worth, A. P. and Fentem, J. H. (1999). A general approach for evaluating stepwise testing strategies. *ATLA 27*, 161-177.

3 Alternatives to skin corrosion/irritation testing in animals

3.1 Background

The Draize method for evaluating skin irritation and/or corrosion has been in existence for a long time and has been used widely to predict the skin irritation/corrosion potential of many chemical compounds (Draize et al., 1944). Rabbits are the species most often used in the assay and the number of animals used in each study can vary from 4 to 12. In many instances, only one dose level is tested. Additionally, irritancy is often assessed as a part of other studies (for example, dermal LD_{50} studies) in which the same animals are used for all assessments. The chemical to be tested may be applied to the abraded or intact skin of the animals' backs that have been shaved or clipped. The site of application is left unoccluded or may be occluded with various materials such as gauze patches and elastic tape. Reactions are usually read according to a scale devised by Draize at intervals ranging from 1 hour to 7 or 14 days.

The relevance of animal irritancy data (most of which has been generated by the Draize assay) to humans is questionable. Moreover, the use of this assay to evaluate corrosivity or irritation has been denounced in recent years for ethical reasons. Numerous alternative methods have been developed to replace the use of animals for evaluating skin corrosion and irritation. These include those based on structure-activity relationships and human patch testing as well as *in vitro* methodologies.

3.2 *In vitro* models

The *in vitro* methodologies which serve as alternatives to skin corrosion/irritation testing in animals range from a "test-tube" assay to excised skin to systems involving culture of skin cells on various mesh-like frameworks. Additionally, the endpoints measured vary from simple assays of dye retention/release to evaluations of more complex variables such as transcutaneous electrical resistance.

The "test tube" assay, also known as Corrositex®, is an *in vitro* method based on the ability of a corrosive chemical or chemical mixture to pass through, by diffusion and/or destruction, a biobarrier and to elicit a color change in the underlying liquid Chemical Detection System (CDS) (ICCVAM, 1999). The biobarrier is composed of a hydrated collagen matrix in a supporting filter membrane, while the CDS is composed of water and pH indicator dyes. Test chemicals and chemical mixtures, including solids and liquids, are applied directly to the biobarrier. The time it takes for a test chemical or chemical mixture to penetrate the biobarrier and produce a color change in the CDS is compared to a classification chart to determine corrosivity/noncorrosivity and to identify the appropriate US Department of Transportation (DOT) packing group. Despite criticism of the assay's ability to discriminate between corrosive and non-corrosive industrial chemicals (for example, Stobbe et al., 1999), the US DOT currently accepts the use of Corrositex® to assign subcategories of corrosivity (packing groups) for labeling purposes according to United Nations (UN) Committee of Experts on the Transport of Dangerous Goods guidelines. However, the US DOT limits the use of Corrositex® to specific chemical classes, including acids, acid derivatives, acyl halides, alkylamines and polyalkylamines, bases, chlorosilanes, metal halides, and oxyhalides. It has also been suggested that Corrositex® may be used as part of a tiered-testing strategy where positive responses require no further testing and negative responses must be followed by dermal irritation testing (Scala et al., 1999).

Another method for assessing corrosivity is the transcutaneous electrical resistance (TER) assay. Human or rat skin is excised and changes in transcutaneous electrical resistance across the *stratum corneum* are evaluated after exposure to a test chemical. This assay has proven to be a good predictor of corrosivity (Botham et al., 1992; Whittle and Basketter, 1993, 1994).

Although, Corrositex® and the TER assay are definitely valuable tools with which to measure the corrosive potential of many chemical compounds, dermal corrosion and irritancy have also been

studied *in vitro* by using various skin irritancy models.

Van de Sandt et al. (1997) classified *in vitro* skin irritancy models into four types: immortalized keratinocyte cell lines, conventional keratinocyte cultures, skin explant or organ cultures, and air-exposed human keratinocyte cultures (epidermal or skin equivalents). They characterized their strengths and weaknesses as follows. Immortalized keratinocyte cell lines (most notably HaCaT cells) closely resemble normal keratinocytes in their growth and differentiation characteristics and respond, *in vitro*, to modulators of differentiation. They exhibit a remarkable stable genetic balance over extended culture periods, making them a favorite model of normal human keratinocytes. They do not possess the differentiation capacity to form an *in vitro* skin equivalent in organotypic cultures, however. Normal human keratinocyte (NHK) cultures offer distinct advantages including flexibility with respect to experimental time, reproducibility, and relative ease of cryopreservation. Their use as submerged monolayer cultures is limited to testing of water-soluble test substances, however, and they lack a *stratum corneum* resulting in increased sensitivity to chemically-induced toxicity. Skin explant and organ cultures closely match the *in vivo* situation when full thickness skin is used. The explant is placed on a grid or insert and incubated at the air-liquid interface to prevent further growth. The resulting brief survival time of the culture limits testing, however, to short exposures. Air-exposed reconstructed cultures represent a fourth *in vitro* skin irritancy model, many versions of which are currently undergoing further development and scrutiny. Essentially, differentiated keratinocyte cultures are grown at the air-liquid interface on various substrates, such as inert filters (Rosdy and Clauss, 1990), collagen sheets (Tinois et al., 1991), de-epidermised dermis (Regnier et al., 1990), and fibroblast-populated collagen gels (Bell et al., 1991). The main characteristics of native skin tissue, including basal cell layer, *stratum spinosum* and *granulosum*, as well as *stratum corneum* are present in these models. The presence of an uninterrupted *stratum corneum* in several *in vitro* models permits the application of water-insoluble compounds and final topical formulations. However, in some models the barrier function is impaired, some models are not available any longer, costs may be high, and results may be variable between batches.

A number of commercial models that incorporate characteristics of the *in vitro* skin irritancy models discussed above have been introduced and studied by various researchers (for example, Living Skin Equivalent™ (Organogenesis), Skin™ (Advanced Tissue Sciences), EpiDerm™ (MatTek), EpiSkin™ (L'Oreal) and PrediSkin™ (BioPredic)) (Fig. 4).

Examples of research investigating skin models and corrosivity include evaluations by Lake et al. (1994), who studied the potential corrosivity of 22 compounds and 17 final product formulations using EpiDerm™. Ninety-seven percent of the compounds evaluated were assigned the proper US DOT corrosive/non-corrosive designation. Additionally, Perkins and Osborne (1994) studied skin corrosion using several commercially available skin models. In their work, human skin equivalent cultures, Models ZK1300™ and ZK1301™ (Advanced Tissue Sciences) and EpiDerm™ were compared using MTT as the endpoint. Corrosive materials were accurately distinguished from strong, moderate, and mild test materials by all three models.

Moreover, numerous comparative studies of the predictive abilities of various models for dermal irritancy have also been conducted (for example, Helman et al., 1992; Roguet et al., 1994; Wolf et al., 1995; Hayden, 1996). Most recently, Roguet and colleagues (1999) made an in-depth effort to evaluate several current models. They found that EpiSkin™, EpiDerm™ a SkinEthic™ model and various "in-house" (L'Oreal) models reproduced many of the characteristics of human epidermis. In particular, histologic examination showed a completely stratified *stratum corneum* in all models. Inter-batch variations were low for EpiDerm™ and moderate for EpiSkin™, but considerable variations (thickness of the epidermis, presence of pycnotic cells) were noted in the SkinEthic™ model. Metabolic studies showed the presence of NADPH quinone reductase and glutathione-S-transferase activities in all commercial models. *In vitro* assessments of skin irritancy were conducted using cytotoxicity (MTT), release of IL-1α, and cytoplasmic enzymes, as endpoints. After SDS treatment, inter-batch variability of MTT results was lower for EpiDerm™, followed by EpiSkin™, the Cosmital™ model and finally the SkinEthic™ model. Results of IL-1α, lactate dehydrogenase and glutamate-oxaloacetate transferase release showed a relatively high variability intra-batch or inter-batch.

Other research efforts have investigated the response of reconstructed skin models when used with high concentrations of test substance and/or with differ-

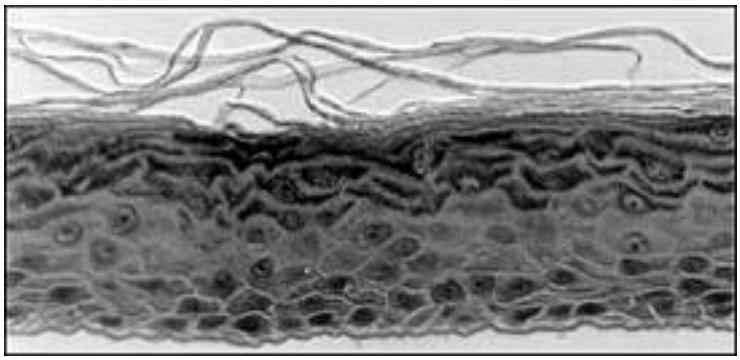

Fig. 4: EpiDerm™ skin model. (Reprinted with permission from MatTek Corporation).

ent product types. Earl and colleagues (1996) questioned the relevance of *in vitro* cell cultures for measuring the cytotoxicity of high concentration test substances, such as surfactants or surfactant mixtures. They compared data generated *in vivo* with that obtained from the agarose overlay cytotoxicity assay, and an MTT time-course assay using EpiDerm™. The results indicated that the agarose overlay assay did not distinguish between any of the treatments; however, the EpiDerm™ assay broadly reflected the results obtained *in vivo* and did distinguish between different surfactants and their mixtures *in vitro*.

Koschier et al. (1997) compared the ability of three-dimensional human skin models to evaluate the dermal irritancy of petroleum products. Three commercially supplied human skin constructs (Living Skin Equivalent™ (LSE), EpiDerm™, and ZK1300™) were treated with 14 petroleum refinery streams. Endpoints measured were lactate dehydrogenase, IL-1α, PGE_2, and MTT conversion. Spearman rank order analysis comparing the *in vitro* cytotoxicity data with the Primary Dermal Irritation Index (PDII) scores gave values of 0.54 (LSE), 0.41 (ZK1300™), and 0.79 (EpiDerm™), respectively. These data indicated that IL-1α concentrations showed reasonable correlations with the known *in vivo* irritation level, especially in the EpiDerm™ cultures. The best prediction of *in vivo* irritation in all three models appeared to come from a combination of cytotoxicity and IL-1α measurements (Fig. 5).

Perkins and co-workers (1999) also investigated the use of human skin equivalent cultures with different product classes. They directly compared *in vitro* to *in vivo* human skin responses using historic or concurrent skin response data for diverse products and ingredients including surfactants, cosmetics, antiperspirants (AP) and deodorants (DO). EpiDerm™ was used and human clinical protocols were paralleled by topical dosing of neat or dilute test substances to the *stratum corneum* surface of the skin cultures. MTT conversion, lactate dehydrogenase and aspartate aminotransferase release and IL-1α expression were monitored. For surfactants, dose-

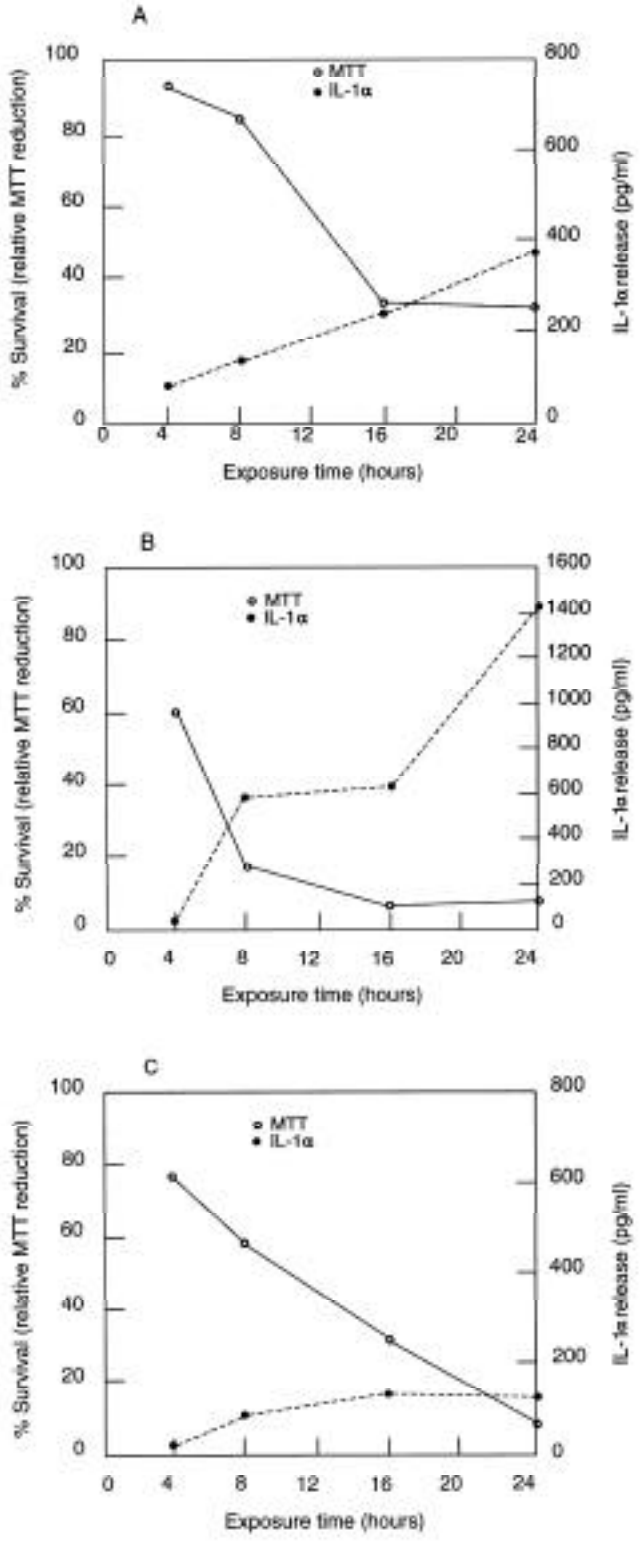

Fig. 5: A comparison of the time course of cytotoxicity (○) estimated by relative MTT reduction, and IL-1α release (●) after direct application of material F-137 to the surface of the three different tissue constructs: (A) LSE, (B) skin² model ZK1300™, and (C) EpiDerm™ (from Koschier et al., 1997).

Tab. 4: Rank-order of surfactant irritancy by *in vivo* and *in vitro* tests (from Perkins et al., 1999)

Surfactant	*In vivo* dose range (µg/cm2)	Surfactant active (% w/w)	*In vivo* rank[a] human skin patch	*In vitro* rank[b] skin cultures
			(Irritancy Groups-A, B, C[c])	
1 Sodium lauryl sulphate	22-430	0.01-0.40	1/A	1/A
2 Anionic ethoxylate (a)	22-430	0.02-0.40	2/B	3/B
3 Anionic ethoxylate (b)	22-430	0.02-0.40	3/B	4/B
4 Anionic ethoxylate (c)	22-430	0.02-0.40	4/B	5/B
5 Nonionic (a)	690-11,000	0.60-10.0	5/C	6/C
6 Amphoteric betaine	207-3300	0.20-3.00	6/C	7/C
7 Nonionic (b)	690-11,000	0.60-10.0	7/C	2/A

Note. Lowercase (a), (b), (c) differentiate surfactants within a class.
[a] *In vivo* rank ordering is based on the cumulative irritation score, where 1 is the most irritating and 7 is the least irritating surfactant.
[b] *In vitro* ranks are based on the dose of surfactant causing 50% cytotoxicity (MTT_{50} values) in human skin cultures, where 1 is the most irritating or cytotoxic, and 7 is the least irritating or cytotoxic surfactant.
[c] Surfactants within the same irritancy group (indicated by letters A, B or C) are not significantly different from each other, but are significantly different ($p<0.05$) from surfactants in other irritancy groups.

response curves of MTT cell viability data clearly distinguished strongly irritating from milder surfactants and rank ordered irritancy potential in a manner similar to *in vivo* patch test results (Tab. 4).

For AP/DO products, all the *in vitro* endpoints correlated well with consumer reported irritation, with IL-1α showing the greatest capacity to distinguish irritancy over a broad range. IL-1α also showed the best prediction of human skin scores from 14-day cumulative irritancy tests of cosmetic products (Fig. 6).

The authors felt that these results confirmed the value of cornified human skin cultures as an *in vitro* preclinical screen for prediction of human skin irritation responses, and identified the need to customize *in vitro* endpoints for irritancy prediction for different product classes.

3.3 Validation and regulatory activities

Several human keratinocyte and human skin models have been evaluated by the European Center for the Validation of Alternative Methods (ECVAM) for their predictive value (Van de Sandt et al., 1997). Furthermore, four alternative models were investigated as potential candidates for corrosion testing by ECVAM in an international validation study. The rat transcutaneous electrical resistance (TER) assay, Corrositex®, Skin^2 ZK1350™ test, and EpiSkin™ were all included in this effort (Fentem et al., 1998).

Two of the four tests evaluated in the validation study, the TER assay and EpiSkin™, met the criteria concerning acceptable underprediction and overprediction rates. Hence, they are considered scientifically validated for use as replacements for the animal test for distinguishing between corrosive and non-corrosive chemicals for all of the chemical types studied. EpiSkin™ was the only test able to distinguish between known R35 (UN packing group I) and R34 (UN packing groups II & III) chemicals, for all of the chemical types included, on an acceptable number of occasions.

The corrosive potentials of about 40% of the test chemicals could not be assessed with Corrositex®, and the assay did not meet all of the criteria for it to be considered acceptable as a replacement test. However, the group did decide that Corrositex® may be valid for testing specific classes of chemicals, such as organic bases and inorganic acids, a decision similar to that reached by the US DOT as discussed above. Subsequent to this large validation effort, a prevalidation exercise has also been conducted with EpiDerm™ re-vealing excellent prediction of corrosivity for a wide spectrum of chemicals (Liebsch, 1999).

The Organization for Economic Cooperation and Development (OECD) has recently developed a tiered-testing strategy for evaluating dermal corrosion and irritation that allows for the use of validated and accepted *in vitro* methods (OECD, 1998). Worth et al. (1998) report an evaluation of the proposed OECD testing strategy as it relates to the classification of skin corrosives. By using a set of 60 chemicals, an assessment was made of the effect of applying three steps in the strategy. The results indicate that chemicals can be classified as corrosive (C) or non-corrosive (NC) with sufficient reliability by the sequential application of three alternative methods, i.e., structure-activity relationships (where available), pH measurements, and a single *in vitro* method (either the rat skin transcutaneous electrical resistance (TER) assay or the EpiSkin™ assay). They concluded that the proposed OECD strategy for skin corrosion can be simplified without compromising its predictivity (for example, it does not appear necessary to measure acid/alkali reserve (buffering capacity) in addition to pH for the classification of pure chemicals).

Following this evaluation of the OECD's tier-testing strategy for corrosives, ECVAM recognized the need to make substantial progress in validating *in vitro* tests for skin irritation (Botham, 1999). A Task Force was established

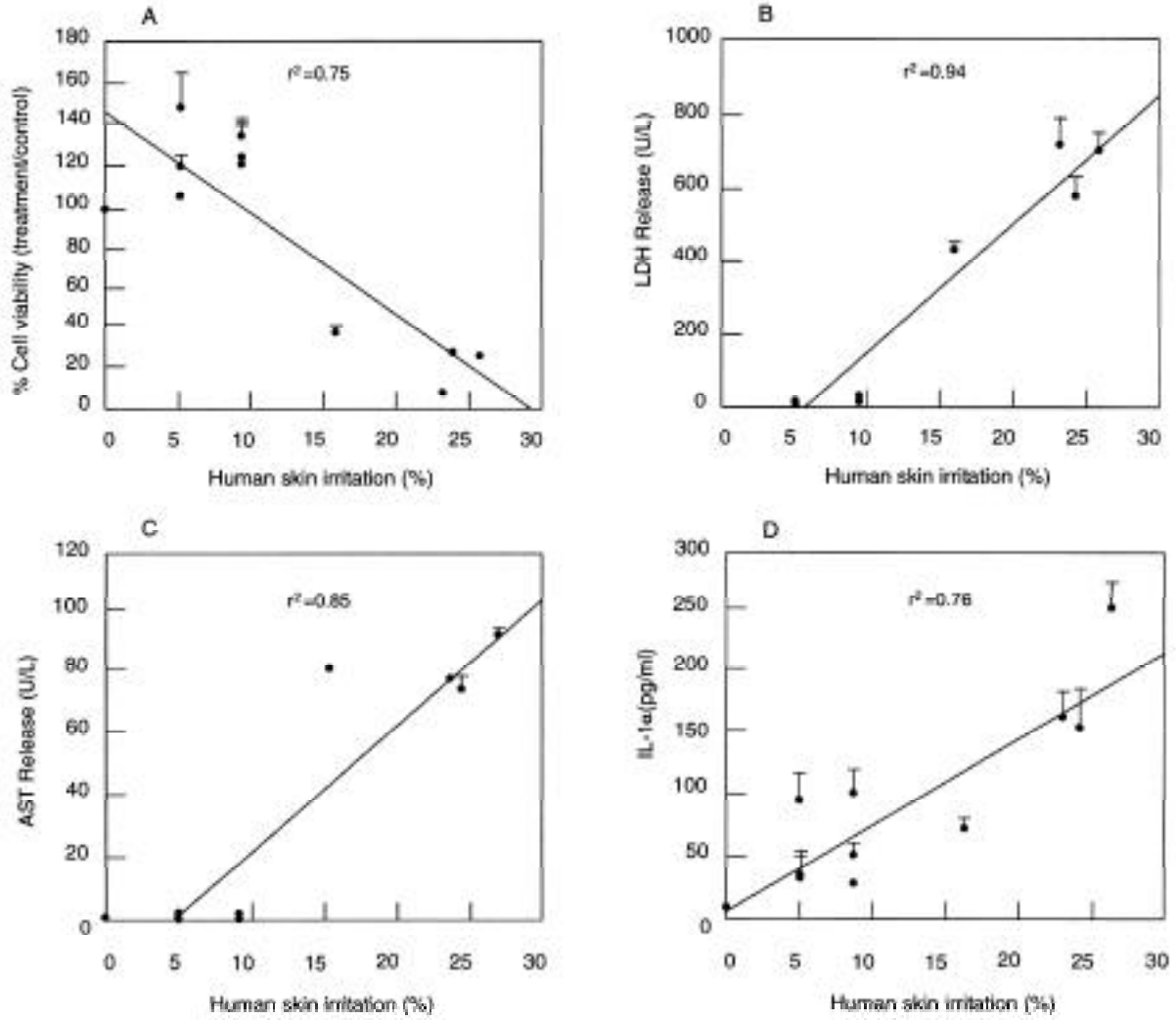

Fig. 6: Antiperspirant/deodorant study 1.
The percent of human subjects reporting either subjective or objective skin irritation for 11 products are plotted vs. each *in vitro* endpoint. The *in vitro* endpoints evaluated include cell viability (MTT assay; a), enzyme release (LDH; b) (AST; c), and extracellular IL-1α (d). Symbols and error bars represent the mean ± SD across four replicate cultures. To compare the *in vitro* endpoints to the human skin irritation % response, a simple linear regression analysis was used. (r^2, 0.75-0.94) (from Perkins et al., 1999).

which issued a challenge to developers/users of relevant methods to submit data, obtained according to a specified protocol with a preliminary prediction model, on ten defined test chemicals. This resulted in selection of four tests – EpiSkin™, EpiDerm™, PrediSkin™, and the non-perfused pig ear model – for prevalidation testing. The main objective of these validation activities is to identify those *in vitro* tests capable of discriminating skin irritants (I) from non-irritants (NI), as defined according to EU risk phrases (R38; no label) and the harmonized OECD criteria (Irritant; no label) (Fentem et al., 1999). Once these tests have completed a prevalidation assessment, a subsequent validation study, if successful, would then allow the complete OECD testing strategy for skin corrosion and irritation to be followed without the use of live animals.

3.4 Summary, conclusions, and future work

Alternative methodologies for the testing of dermally corrosive or irritating substances include a "test-tube" assay, measurements of transcutaneous electrical resistance in excised skin, and *in vitro* skin culture models of several types. Structure-activity-relationship modeling as well as pH determinations and acid/alkali reserve evaluations may also contribute to the successful determination of dermal corrosivity or irritancy of a test compound, particularly if used in tier-testing strategies. Of the *in vitro* skin irritancy systems available, skin model kits appear to be the most popular candidates, perhaps due to the ease with which they can be used. Although the variabili-

ty of their response is still under investigation, it appears to be due in many cases to factors such as test compound concentration and product type.

Regulatory acceptance of *in vitro* models of skin corrosion or irritation appears to be no longer as substantive a hurdle as it once was as evidenced by recent US DOT and OECD initiatives. Models currently validated for use as alternatives to corrosivity testing in animals include Corrositex® (limited to specific classes of chemicals), the rat TER assay, and EpiSkin™. EpiDerm™ has also been prevalidated as an alternative method to corrosion testing. Validation of *in vitro* methodologies for skin irritation appears to be the major "next step" in development of alternatives acceptable to regulators. Models of dermal irritancy currently undergoing this process include EpiSkin™, EpiDerm™, Prediskin™, and the nonperfused pig ear model (Elliott et al., 1999) with results expected by the end of 2000.

The use of alternatives to dermal corrosivity or irritancy testing in animals will advance greatly the aims of the alternatives to animal testing movement for several reasons. The first, and perhaps, foremost advantage to use of these methods is the abolishment of testing corrosive and/or irritating substances on live animals. The humane implications of this action are straightforward and it clearly represents appropriate implementation of a replacement alternative. Secondly, the number of animals used will be reduced by use of these assays (for example, Corrositex® uses no animal cells at all, a small section of rat skin is used in the TER assay, and most *in vitro* skin irritancy models incorporate human skin). Finally, from a scientific perspective, a test which has been questioned not only with regard to its reliance on subjective observations, but also as to its more general relevance to human skin corrosion and irritation, will have been replaced, in many instances, with *in vitro* assays using more objectively evaluated endpoints. The challenge remains however, to relate the endpoints measured in these assays to the development of dermal corrosion and/or irritation by more than superficial explanations.

The next five years of effort in this area of alternatives to animal testing should result in widespread acceptance and integration of tier-testing strategies for corrosivity testing. Additionally, validation and regulatory acceptance of *in vitro* assays of dermal irritancy should also be accomplished. The next ten years of research should result in refined *in vitro* skin irritancy models exhibiting very little variability in response and in much greater use in testing laboratories. Moreover, the relationship between the endpoints measured in these assays and production of dermal irritancy should be better understood. The next 20 years of research should yield models that will effectively discriminate between different types and levels of dermal insult perhaps lending insight into the mechanisms lying beneath the subsequent pathological response. Progenitors of this type of "specialized" model include Melanoderm™ and Epi-201™ (both from MatTek). Melanoderm™ is a system into which melanocytes have been incorporated that may be useful for studying melanogenesis, and skin pigmentation (Klausner et al., 1999). Epi201™ is a stratified culture of proliferating and differentiating human keratinocytes which, when treated with epidermal growth factor "develops a granular layer which is thicker and more heavily populated with keratohyalin granules compared to standard EpiDerm™." Studies of re-epithelialisation, wound healing and normal barrier development are thought to benefit from use of models such as this one (Klausner et al., 1998).

References

Bell, E., Parenteau, N., Gay, R. et al. (1991). The living skin equivalent: its manufacture, its organotypic properties and its response to irritants. *Toxic. In Vitro 5*, 519-596.

Botham, P., Hall, T., Dennet, R. et al. (1992). The skin corrosivity test in vitro. Results of an inter-laboratory trial. *Toxic. In Vitro 6*, 191-194.

Botham, P. (1999). Skin corrosion and irritation: Toward complete replacement. *ATLA 27*, 108.

Draize, J., Woodward, G. and Calvery, H. (1944). Methods for the study of irritation and toxicity of substances applied topicaly to the skin and mucous membranes. *J. Pharm. and Exper. Therap. 82*, 377-390.

Earl, L., Hall-Manning, T., Holland, G. et al. (1996). Skin irritation potential of surfactant mixtures: using relevant doses in in vitro systems. *ATLA 24*, 249. Abstract #73.

Elliott, G., Pool, C. and Bruijnzeel, P. (1999). Isolated pig ears: An in vitro model for screening for irritant chemicals using stimulation of transepidermal water loss as the test parameter. *ATLA 27*, 110.

Fentem, J., Archer, G., Balls, M. et al. (1998). The ECVAM international validation study on in vitro tests for skin corrosivity. 2. Results and evaluation by the management team. *Toxic. in Vitro 12*, 483-524.

Fentem, J., Botham, P., Harbell, J. et al. (1999). Prevalidation of in vitro tests for acute skin irritation. *ATLA 27*, 346.

Hayden, P. (1996). Cytokine and lipid mediator release from in vitro skin models following detergent treatment or physical wounding. *J. Invest. Dermatol. 106(4)*, 936. Abstract #782.

Helman, M., Biesaga, S. and Koslo, R. (1992). Reformed human skin: Potential screening tool for dermal irritation testing. *The Toxicologist 12(1)*, 297. Abstract #1147.

ICCVAM (Interagency Coordinating Committee for the Validation of Alternative Methods) (1999). Corrositex®: An in vitro test method for assessing dermal corrosivity potential of chemicals. NIH Publication No. 99-4495. June.

Klausner, M., Neal, P. and Kubilus, J. (1998). Initial results with Epi-201, a developing in vitro model of human epidermis. *J. Invest. Dermatol. 110(4)*, 558. Abstract. #513.

Klausner, M., Neal, P., Breyfogle, B. and Kubilus, J. (1999). Melanoderm: an epidermal tissue model containing functional melanocytes. *ATLA 27*, 299.

Koschier, F., Roth, R., Wallace, K. et al. (1997). A comparison of three-dimensional human skin models to evaluate the dermal irritation of selected petroleum products. *In Vitro Toxicol. 10(4)*, 391-405.

Lake, L., Swanson, J., Kubilus, J. et al. (1994). Use of an in vitro skin model to predict chemical corrosivity. *The Toxicologist 14(1)*, Abstract #340.

Liebsch, M., Traue, D., Barrabas, C. et al. (1999). Prevalidation of the EpiDerm™ skin corrosivity test. *ATLA 27*, 350.

OECD (Organization for Economic Cooperation and Development) (1998). Harmonized Integrated Hazard Classification System for Human Health and Environmental Effects of Chemical Substances. Paris: OECD.

Perkins, M. and Osborne, R. (1994). Application of an in vitro method for skin corrosion testing to a variety of commercially available skin cultures. *J. Invest. Dermatol. 102(4)*, 641. Abstract # 707.

Perkins, M. A., Osborne, R., Rana, F. R. et al. (1999). Comparison of in vitro and in vivo human skin responses to consumer products and ingredients with a range of irritancy potential. *Toxicol. Sci. 48(2)*, 218-229.

Roguet, R., Regnier, M., Cohen, C. et al. (1994). The use of in vitro reconstituted human skin in dermotoxicity testing. *Toxic. In Vitro 8(4)*, 635-639.

Roguet, R., Beck, H., Boelsma, E. et al. (1999). Testing and improvement of reconstructed skin kits in order to elaborate European standards: first results. *ATLA 27*, 333.

Regnier, M., Asselineau, D. and Lenoir, M. (1990). Human epidermis reconstructed on dermal substrates in vitro: an alternative to animals in skin pharmacology. *Skin Pharmacology 3*, 70-85.

Rosdy, M. and Clauss, L. (1990). Complete human epidermal cell differentiation in chemically defined medium at the air-liquid interface on inert filter substrates. *J. Invest. Dermatol. 95*, 409-414.

Scala, R., Stokes, W., Hill, R. et al. (1999). The results of an independent peer review panel evaluation of the validation status of Corrositex®. *ATLA 27*, 334.

Stobbe, J., Drake, K. and Maier, K. (1999). Comparison of dermal corrosion values for selected industrial chemicals using Corrositex®. *The Toxicologist 48(1S)*, 77. Abstract #359.

Tinois, E., Tiollier, J., Gaucherand, M. et al. (1991). In vitro and post-transplantation differentiation of human keratinocytes grown on human type IV collagen film of a bilayered dermal substitute. *Exper. Cell Res. 193*, 310-331.

Van de Sandt, J., Roguet, R., Cohen, C. et al. (1997). The use of human keratinocytes and human skin models for predicting skin irritation. The report and recommendations of ECVAM workshop 38. *ATLA 27*, 723-743.

Whittle, E. and Basketter, D. (1993). The in vitro skin corrosivity test. Development of method using human skin. *Toxic. In Vitro 7(3)*, 265-268.

Whittle, E. and Basketter, D. (1994). In vitro skin corrosivity test using human skin. *Toxic. In Vitro 8(4)*, 861-863.

Wolf, B., Kwan, S., Moral, J. and Rheins, L. (1995). Development of in vitro methods for evaluating the epidermal and dermal effects of alpha hydroxy and amides. *J. Invest. Dermatol. 104(4)*, 639. Abstract #511.

Worth, A. P., Fentem, J., Balls, M. et al. (1998). An evaluation of the proposed OECD testing strategy for skin corrosion. *ATLA 26(5)*, 709-720.

4 Alternatives to skin sensitization testing in animals

4.1 Background

De Silva et al. (1996) described the relationship between skin sensitization and *in vitro* alternatives as follows: "The induction of contact hypersensitivity involves a complex series of cellular and molecular events. The relative contributions and interdependence of these events are not completely understood. It remains to be demonstrated whether an *in vitro* test, or several *in vitro* tests in combination which model the critical steps in sensitization, can replace animal experiments for predicting contact allergic reactions in humans." Hence, the intricate cascade of events leading to development of skin sensitization, as elegant as it is, represents a substantial hurdle to the development of *in vitro* assays not unlike the development of assays for a similarly intricate biological system, that of reproduction. Alternatively, the sensitization "cascade" also provides fertile ground for exploration of novel assay techniques that do not use animals.

Historically, the guinea pig has been used extensively as the animal model of choice for sensitization testing. Testing techniques vary, but the basic protocol scheme involves initial application of the sensitizer to the skin of the animal, followed by a delayed "challenge" exposure to the same sensitizer after a given interval of time. Positive reactions are scored according to degree of redness and/or swelling at the site of application. The test can be quite lengthy taking anywhere from 3 to 6 weeks to complete.

Current regulatory guidelines include several methods of skin sensitization testing. The guinea pig maximization test (GPMT) (Magnusson and Kligman, 1970) and Buehler assay (BA) (Buehler, 1965) seem to be the most widely used techniques, endorsed by the EU, OECD, and the United States (EU, 1992; OECD, 1993; USEPA, 1998). USEPA (United States Environmental Protection Agency) Guidelines suggest that, "Either sex may be used in the Buehler test and the GPMT. If females are used, they should be nulliparous and not pregnant. The Buehler test recommends using a minimum of 20 animals in the treatment and at least 10 as controls. At least 10 animals in the treatment group and 5 in the control group should be used with the GPMT, with the stipulation that if it is not possible to conclude that the test substance is a sensitizer after using fewer than 20 test and 10 control guinea pigs, the testing of additional animals to give a total of at least 20 test and 10

control animals is strongly recommended." Moreover, these guidelines state that, "The dose level will depend on the test method selected. In the Buehler test, the concentration of the induction dose should be high enough to cause mild irritation, and the challenge dose should use the highest non-irritating concentration. In the GPMT, the concentration of the induction dose should be well tolerated systemically, and should be high enough to cause mild-to-moderate skin irritation; the GPMT challenge dose should use the highest non-irritating concentration."

These requirements result in the use of thirty animals per test, at minimum, which is a relatively large number when compared to the number of animals used in several of the alternative assays which are discussed below. Furthermore, these techniques can employ an adjuvant that may enhance the severity of the response observed in test animals. Finally, as with the Draize eye and skin irritation tests, the grading of responses in either of the sensitization assays discussed above depends upon the subjective, though often well-trained, assessment of a human observer.

4.2 Alternative assays

Each of the alternative assays discussed below attempts to capture data representing a particular step or steps in the "cascade" of events leading to contact sensitization. Because of the many steps involved, numerous opportunities for the capture of such data exist. However, techniques for "hard-to-catch" events are still evolving.

Two of the more well-known alternative assays are the mouse ear-swelling test (MEST) and the local lymph node assay (LLNA). Both assays are used as screening tests for skin sensitization. The LLNA has also recently been evaluated for use as a "stand-alone" assay of skin sensitization (ICCVAM-NICEATM, 1999).

The MEST measures increases in ear-thickness (swelling) in the mouse that result from applying a test chemical to the animal's ear after these animals have been exposed to the test material in an earlier, induction phase of the protocol (Gad et al., 1986). Measurement of ear-thickness yields quantitative data giving the MEST a decided advantage over the more subjective observations used in guinea pig tests of sensitization. The MEST has, however, been criticized for its lack of sensitivity (Cornacoff et al., 1988; Dunn et al., 1990). Subsequent research has indicated that this lack may be remedied, at least in part, by diets rich in Vitamin A (Thorne et al., 1994; Sailstad et al., 1993).

The local lymph node assay measures events occurring during the induction phase of the sensitization reaction. Specifically, this assay most often measures the uptake of tritiated methyl thymidine by proliferating T-cells in draining lymph nodes of mice exposed to a test chemical (Kimber and Basketter, 1992). The LLNA has been studied and compared extensively with other assays of skin sensitization (for example, Kimber et al., 1990, 1991, 1995; Hayes et al., 1998; Hayes and Meade, 1999; Ohta et al., 1999; Howell et al., 2000). Moreover, an evaluation of skin sensitization methods by ECVAM suggested that the assay "offers a number of advantages compared with guinea pig methods, not least of which are that the LLNA is relatively rapid and cost-effective. The endpoint is quantitative and the assay does not require the use of an adjuvant" (De Silva et al., 1996). The recent peer review of the LLNA conducted by the Interagency Coordinating Committee on Validation of Alternative Methods and the National Interagency Committee for the Evaluation of Alternative Methods (ICCVAM-NICEATM) came to similar conclusions. They stated that, "the assay performed at least as well as currently accepted guinea pig methods (GPMT/BA) for the hazard identification of strong to moderate chemical sensitizing agents" (ICCVAM-NICEATM, 1999). They also agreed with ECVAM that the assay has numerous aspects that make it a favorable alternative assay (for example, reduces animal distress, potentially reduces animal numbers). The LLNA has demonstrated several weaknesses common to alternative assays of skin sensitization (for example, failure to detect weak sensitizers, identification of strong irritants as sensitizers). Researchers in this field feel, however, that these irregularities may be explained with further in-depth study.

Additional, suggested alternative approaches to the study of skin sensitization have included study of the quantitative structure-activity relationships of haptens, cytokine response induced in cultured keratinocytes by haptens, contact allergen-induced alterations in adhesion molecules on keratinocytes and epidermal Langerhans cells, and hapten- mediated T-cell activation by Langerhans cells (Elmets, 1996). Evaluation of various cell cultures systems (e.g. dendritic cells, keratinocytes, skin explants, and co-cultures of skin cells) that support several of these approaches has been made as well (De Silva et al., 1996).

One such approach involves interleukin 1ß (IL-1ß) production by Langerhans cells. This event has become linked to development of an alternative assay for skin sensitization because it has been shown to be a primary mediator of the induction phase of contact sensitivity (Enk and Katz, 1992; Enk et al., 1993). Development of this assay has been hindered by the difficulty of obtaining Langerhans cells. However, several studies seem to indicate that cultures of dendritic cells isolated from peripheral blood can be utilized effectively to study contact allergens (Romani et al., 1994; Lenz et al., 1993; Hoppe et al., 1995).

Reutter and colleagues (1997) studied early upregulation of IL-1ß in dendritic cells (DCs) following exposure to contact sensitizers. DCs were cultivated in serum-free medium supplemented with granulocyte/macrophage-colony stimulation factor (GM-CSF) and Interleukin 4 (Il-4). The DCs exhibited typical dendritic morphology, characteristic surface markers and high stimulatory capacity for T-cells. Five-day-old DCs were incubated with contact sensitizers, pentadecyl-catechol, 2,4,6-trinitrobenzene sulfonic acid, 2,4-dinitrofluorobenzene, $NiSO_4$, $K_2Cr_2O_7$ and the irritant, sodium dodecyl sulfate (SDS). IL-1ß mRNA expression was monitored by using a reverse transcriptase-polymerase chain reaction technique and non-radioactive hybridization procedures. For all contact sensitizers, expression of IL-1ß mRNA increased, whereas treatment with the irritant, SDS had no significant effect on IL-1ß expression (Fig. 7).

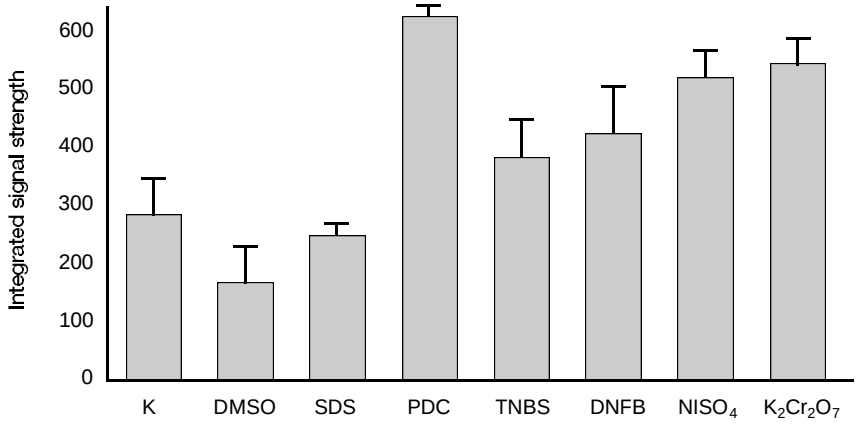

Fig. 7: IL-1ß mRNA expression induced by SDS, PDC, TNBS, DNFB, nickel sulfate and potassium dichromate.
Signal strength was assessed by densitrometric scanning of autoradiograms and integration by WINCAM 1.0 Software. IL-1ß data were standardized to the respective ß-actin controls. Data were expressed in arbitrary units and are representative of three experiments. K = control for probes solubilized in medium; DMSO = control for probes solubilized in DMSO (from Reutter et al., 1997).

In a subsequent study, Pichowski and co-workers (2000) evaluated further the findings of Reutter using dendritic cells obtained from human blood. These cells were cultured in the presence of interleukin-4 (IL-4) and granulocyte/macrophage colony stimulating factor. After 5 days in culture, they exhibited a Langerhans cell-like phenotype. These DC were then cultured with the contact allergen, 2,4-dinitrofluorobenzene (DNFB), SLS or vehicle alone and mRNA expression for IL-1ß mRNA (two-to threefold) in cells derived from three out of eight donors whereas IL-6 and IL-18 were largely unaffected by allergen exposure. Sodium dodecyl sulfate treatment did not induce IL-1ß mRNA expression in any of the donors investigated. Cytokine mRNA expression was also analyzed using the protocol described by Reutter. This protocol did not enhance the sensitivity of measurement of induced cytokine expression. These researchers concluded that, "Although selected upregulation of IL-1ß by blood derived DC has been confirmed, a wider range of contact allergens and irritants needs to be assessed before this approach could be considered for hazard identification".

De Silva and colleagues (1996) describe the migration of hapten-modified Langerhans cells from the epidermis to the regional lymph nodes where they stimulate hapten-specific T-cells as another measurable endpoint of contact sensitization. They cite several studies in which human skin explants have been used in assays for this purpose with reasonable success (for example, reconstituted basement membrane (Kobayashi et al., 1994), reconstituted human epidermis (Tammi and Maibach, 1987; Messadi et al., 1988; Le Poole et al., 1991)).

Additionally, the activation and proliferation of antigen-specific T-lymphocytes caused by primary contact of T-cells with antigen complexes expressed on antigen presenting cells (APC) has been investigated as a measure of contact sensitization. Moulon and associates (1993) examined the capacity of human Langerhans' cells (LC) to sensitize autologous T-cells to the trinitrophenyl hapten (TNP) *in vitro*. They found that two-day cultured Langerhans' cells, but not freshly prepared Langerhans' cells, can induce *in vitro* primary proliferative reactions to the TNP hapten (Fig. 8).

The primary *in vitro* response was strongly inhibited by monoclonal antibodies to major histocompatibility complex (MHC) class I and II, CD4 antigen, and ICAM-1 and LFA-3 adhesion molecules. These researchers concluded that "this *in vitro* priming assay, using cultured human Langerhans' cells as APC, might be useful to analyze the early steps of T-cell sensitization and subsequently to develop *in vitro* predictive tests allowing detection of sensitizing compounds".

Additional work by Krasteva et al. (1996) showed that the two-day cultured

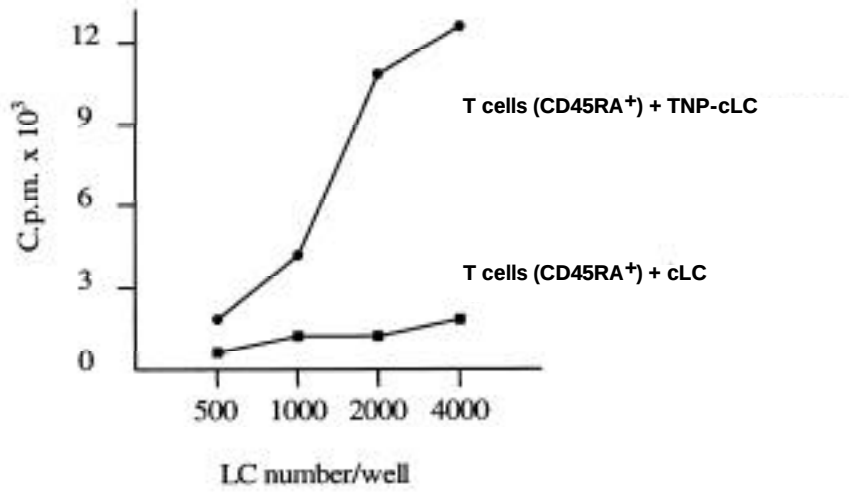

Fig. 8: Naive T-cell sensitization to TNP by cultured LC
$CD45RA^+$ T lymphocytes (10^5) were cultured for 5 days with graded numbers of either non-modified or TNP-treated 2-day incubated LC. T-cell proliferation was assessed by a (H^3)thymidine pulse for the final 18 hrs of culture (from Moulon et al., 1993).

Tab. 5: Forty-eight hour incubation of eight-day-old Langerhans-like dendritic cells (LLDC) with hapten is necessary to induce a proliferative response of autologous T lymphocytes (from Rougier et al., 2000).

	Without incubation		48h incubation	
Hapten	*SI*	Positive experiment	*SI*	Positive experiment
SDS	1.1 ± 0.04	0 (8)	1.7 ± 0.20	0 (8)
TNP	1.9 ± 0.06	0 (5)	4.3 ± 0.68	5 (5)
FITC	1.2/1.8	0 (2)	3.0/3.5	2 (2)
BB	1.5 ± 0.21	0 (8)	5.0 ± 0.56	8 (8)
pPDA	1.1 ± 0.12	0 (8)	2.1 ± 0.57	2 (8)
Coumarin	1.1 ± 0.23	0 (8)	2.1 ± 0.32	1 (8)
Citronellal	1.3 ± 0.06	0 (8)	2.4 ± 0.41	1 (8)

Stimulation index (*SI*) is expressed as mean index values ± 2 S.E.. Positive experiments refer to the number of experiments with an *SI* above 3; the number of cultures effected with different donors is indicated in brackets.

human Langerhans'cell model discriminated between two groups of substances: strong contact sensitizers (TNP, fluorescein isothiocyanate (FITC), Bandrowski's base (BB) and weak sensitizers (coumarin, citronellal, and hydroxycitronellal) and irritants (SDS). They suggested that the assay could be used as a screening *in vitro* assay to eliminate strong contact allergens before further predictive animal tests have to be performed.

Rougier et al. (2000) have refined this work further by studying Langerhans-like dendritic cells (LLDC) derived from cord blood progenitors and autologous T-lymphocytes, isolated from the same blood sample. Treatment of day 12-14 LLDC, with strong haptens trinitrobenzene sulfonic acid (TNP), fluorescein isothiocyanate (FITC) or Bandrowski's base (BB), resulted in the proliferation of T-lymphocytes, whereas weak allergens and irritants, such as sodium dodecyl sulfate (SDS) were ineffective (Tab. 5).

The use of immature (day 8) LLDC and the addition of a 48 hour stage of incubation after hapten contact, resulted in phenotypic maturation of LLDC in addition to lymphocyte activation in all the cultures with strong haptens. The 48 hour stage of incubation, results in sensitization and in some cases the induction of T-cell proliferation to citronellal (1/8), coumarin (1/8) and to a prohapten p-phenylenediamine (pPDA); 2/8). The phenotype of DC after 48 hours of contact with a strong hapten, becomes that of a mature DC (CD83+, CD86+ and HLA-DR++). With fragrance molecules, weak haptens and prohaptens, a comparable phenotype is observed only when T-lymphocytes are activated. These authors suggest that the unresponsiveness observed with weak haptens, may be the consequence of an incomplete maturation of LLDC.

Hence, although further work is undoubtedly needed, three distinct events in the "cascade" leading to contact sensitization have been translated into the alternative assays discussed above: upregulation of IL-1ß production by Langerhans cells, migration of hapten-modified Langerhans cells to the lymph node, and the induction of T-cell proliferation by hapten-modified Langerhans cells.

Cytokines have also been the focus of efforts directed toward development of alternatives to contact sensitization. Although early work with keratinocyte cultures investigated the role of proinflammatory key mediators as possible

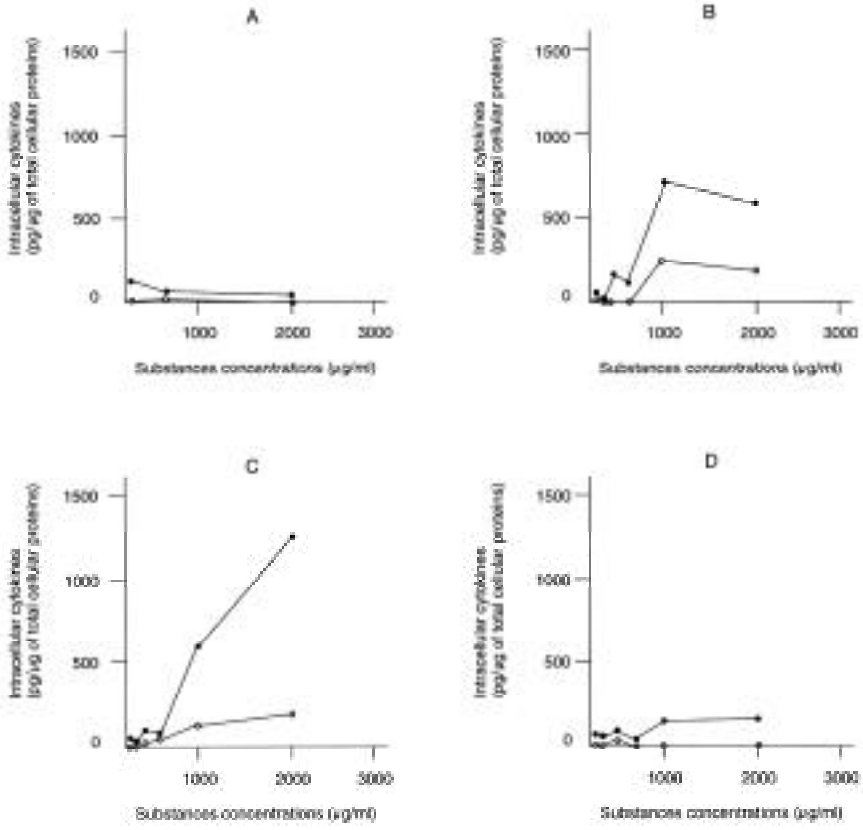

Fig. 9: Influence of increasing doses of Tween 80 (A), Triton X100 (B), BC (C) and DNCB (D) on the IL-1α (•) and IL-8 (∘) intracellular level.
The tissues were exposed to the different concentrations of the tested products for 20 hrs at 37°C (5% CO_2) at which time IL-1α and IL-8 were extracted and quantified by ELISA. The results of one representative experiment are given and expressed in pg of cytokines/µg protein measured in the extract (from Coquette et al., 1999).

alternatives to the Draize skin irritation assay (for example, Muller-Decker et al., 1994), recent studies have incorporated *in vitro* methods that measure the kinetics of cytokine production in response to allergen exposure. Doubts have been expressed, however, as to the discriminatory power of this type of assay because many cutaneous chemical allergens are also primary irritants (De Silva et al., 1996).

A recent example of the combined use of *in vitro* culture of skin cells and measurement of cytokine production after contact allergen exposure is that of Coquette and co-workers (1999). They evaluated the differential expression and release of cytokines by an *in vitro* reconstructed human epidermis following exposure to skin irritant and sensitizing chemicals. A model of reconstructed human epidermis (RHE) was used as an *in vitro* model to discriminate the effects of Tween 80, Triton X100 and benzalkonium chloride (BC) as irritants and 1-chloro-2,4-dinitrobenzene (DNCB) as sensitizing agent. Epidermal irritation and sensitization were evaluated morphologically and on the basis of intracellular and extracellular levels of interleukin-1α and interleukin-8 (IL-8). The corresponding constitutive mRNA levels were quantified and the cytotoxic response was assessed by MTT assay. Topical application of the three surfactants resulted in significant changes of tissue morphology and was coupled with different dose-dependent decreases in cell viability corresponding to their *in vivo* irritant potency. The IL-1α release increased concomitantly with cell viability decrease, but surprisingly, the surfactants did not elicit elevated IL-8 levels. In contrast, DNCB did not induce elevated IL-1α release although it induced a rapid dose-dependent decrease in cell viability. DNCB did increase IL-8 release. Polymerase chain reaction analysis demonstrated the presence of mRNA for IL-1α and IL-8; IL-1α was the most abundant transcript found. BC, Triton X100 and DNCB upregulated IL-8 mRNA expression. These researchers concluded that the production of IL-1α and its release into the extracellular medium was ascribed not only to direct cytotoxicity but also to the extent of tissue stimulation. Their results demonstrated divergent IL-1α and IL-8 release profiles and corresponding mRNA upregulation in response to irritants and DNCB (Fig. 9 and 10, respectively).

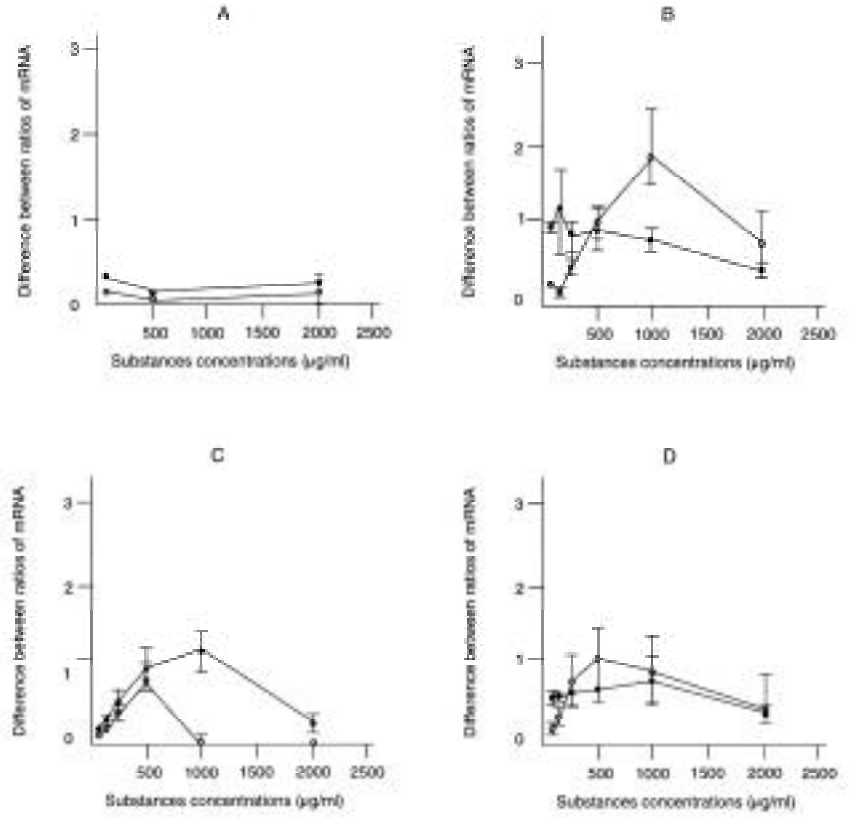

Fig. 10: Influence of increasing doses of Tween 80 (A), Triton X100 (B), BC (C) and DNCB (D) on the IL-1α (•) and IL-8 (∘) mRNA expression in the RHE.
The tissues were exposed to the different concentrations of the tested products for 20 hrs at 37°C (5% CO_2) at which time specific mRNA were extracted and quantified by RT-PCR. Each value represents the mean ± SD of three independent experiments and is expressed as the difference between the ratio of cytokine mRNA/1B15 mRNA determined in treated samples and the same ratio determined in untreated samples (from Coquette et al., 1999).

Although additional testing is needed to clearly delineate irritant and DNCB response, it was suggested that it may be possible to use a single assay that incorporates a reconstituted skin model with measurements of cytokine expression to distinguish between irritant and sensitizing agents as a function of induced cytokine production patterns and cell viability measurements.

Cytokines have also been proposed as additional (non-radioactive) endpoints to be used in conjunction with the local lymph node assay, specifically to detect moderate sensitizers, such as mercaptobenzothiazole and hexyl cinnamic aldehyde (Hariya et al., 1999). This assay measures corrected IL-2 index (CII) defined as including corrections for lymph node weight ratio and ratio of CD4-positive subset. The authors concluded that measurement of CII significantly increased the sensitivity of the LLNA to include detection of moderate sensitizers.

Dearman and colleagues (1999) have also investigated cytokine endpoints for the LLNA: interferon-gamma and interleukin 12. Production of these cytokines by draining lymph nodes was investigated as an alternative "readout" for the LLNA. Animals were exposed to a range of skin sensitizers at two application concentrations. Secretion of IFN-gamma and the p40 subunit of IL-12 by draining lymph node cells was measured by cytokine-specific enzyme-linked im-

munosorbent assay. Parallel experiments were run in which lymph node cell activity was assessed by the usual method of tritiated thymidine incorporation *in situ*. All allergens produced proliferative responses. Treatment with chemical allergen in each case caused a marked increase in IFN-gamma secretion, with particularly high production following exposure to oxazolone or hexyl cinnamic aldehyde. In contrast, application of chemical allergens was not generally associated with elevated IL-12 p40 secretion. However, cytokine secretion did not correlate closely with induction of lymph node cell proliferation. These researchers concluded that expression of cytokines IFN-gamma or IL-12 by allergen-activated lymph node cells does not provide a reliable or sensitive endpoint for the LLNA compared to tritiated thymidine incorporation *in situ*.

Several investigations have been performed in order to define the cytokine signals that are needed by Langerhans cells in order to migrate to regional lymph nodes after exposure to a contact allergen. Cumberbatch et al. (1997) found that both tumor necrosis factor-alpha and interleukin-1ß are needed by Langerhans cells for such migration. They suggest that the changes induced in LC by TNF-α and IL-1ß may include the altered expression of adhesion molecules and acquisition of the ability to interact with and pass through the basement membrane. Subsequent work has suggested that dexamethasone, a synthetic glucocorticoid and immunosuppressant, regulates the *de novo* synthesis and/or release of TNF-α that inhibits LC migration to regional lymph nodes (Cumberbath et al., 1999).

Hence, cytokines continue to be researched extensively as integral components of the contact sensitization response. They may perhaps be used effectively as measurable endpoints in alternative assays pending further development of information about their role in the process of sensitization.

4.3 Validation and regulatory activities

Incorporation of alternative assays into assessments of contact sensitization has been addressed by several regulatory initiatives. Both the EU and OECD describe structure-activity relationship development, the MEST and the LLNA as acceptable screening procedures in assessments of contact sensitization (EU, 1992; OECD, 1993). If the screening test gives a positive result, the substance is regarded as a skin sensitizer labeled with the EU risk phrase R43 indicating "may cause sensitization by skin contact". Negative results obtained in screening tests have to be confirmed by using one of the *in vivo* guinea pig tests described above.

Furthermore, US regulatory agencies acting through the Interagency Coordinating Committee for Alternative Methods (ICCVAM) and National Interagency Committee for the Evaluation of Alternative Methods (NICEATM) have evaluated the LLNA as a "stand-alone" assay of contact sensitization. The review "involved the evaluation of data on 209 chemicals, of which both LLNA and guinea pig data were available for 126 chemicals and both LLNA and human (HMT and HPTA) data were provided for 74 chemicals. An in-depth review of all the chemicals that have been defined in the published literature as human allergens was not conducted for this evaluation. From the analysis generated during the review process, the accuracy of the LLNA vs. GPMT/BA was 89% (N=97), LLNA vs. all guinea pig tests (GPT) was 86% (N=126), the LLNA vs. human data was 72% (N=74), GPMT/BA vs. human was 72% (N=57), and all guinea pig tests (GPT) vs. human was 73% (N=62). In terms of accuracy, sensitivity, specificity, and positive and negative predictivity, the PRP found the performance of the LLNA to be similar to that of the GPMT/BA. Equally important, the performance of the LLNA and the GPMT/BA was similar when each was compared to human data (HMT/HPTA)" (ICCVAM-NICEATM, 1999).

Moreover, the ICCVAM-NICEATM peer review panel (PRP) unanimously recommended the LLNA as a stand-alone alternative for contact sensitization hazard assessment, provided that some protocol modifications were made (for example, lymphocyte proliferation data should be collected at the level of the individual animal). The PRP also concluded unanimously, "that the LLNA is a definite improvement with respect to animal welfare (i.e., refinement and reduction) over the currently accepted GPMT".

In an additional independent assessment of the LLNA in which the ICCVAM-NICEATM report was reviewed as well as publications published since the report and reports from other agencies, the ECVAM Science Advisory Committee also unanimously endorsed the LLNA. They stated that, "the LLNA is a scientifically validated test which can be used to assess the skin sensitization potential of chemicals. The LLNA should be the preferred method, as it uses fewer animals and causes less pain and distress than the conventional guinea-pig methods. In some instances, and for scientific reasons, the conventional methods can be used" (ECVAM, 2000).

Hence, the current regulatory posture embraces alternative methods for evaluation of contact sensitization, either as screening procedures or as "stand-alone" assays.

4.4 Summary, conclusions, and future work

Alternative assays exist for the assessment of contact sensitization. Several are more well-validated than others, but all address measurable endpoints representative of contact sensitization. The LLNA and MEST are both considered appropriate as screening procedures. Moreover, the LLNA has also been endorsed as a "stand-alone" procedure for assay of sensitization. Additional assays are also under development, all of which measure a particular event or series of events leading to sensitization. Examples include interleukin 1ß (IL-1ß) production by Langerhans cells, migration of hapten-modified Langerhans cells from the epidermis to regional lymph nodes, and the activation and proliferation of antigen-specific T-lymphocytes caused by primary contact of T-cells with antigen complexes expressed on antigen presenting cells (APC). Cytokine profiles are also under scrutiny as possible markers of sensitization response.

Progress toward the development and refinement of additional alternative assays of contact sensitization is strongly

dependent upon breakthroughs in our understanding of the immune processes mediating the response. In fact, one might say that they have and will continue to proceed "hand-in-hand". The extensive efforts directed toward validation of the local lymph node assay have borne the much-needed fruit of a "stand-alone" assay that incorporates elements of both refinement and reduction. This type of effort is needed to produce additional, validated assays of contact sensitization, particularly replacement assays. These efforts should be encouraged by the regulatory initiatives already in place. However, much more basic research remains to be done before a fully validated replacement assay of contact sensitization finds regulatory support. One of the most pressing issues at present seems to be the ability of any potential replacement assay to adequately discriminate between processes leading to dermal irritation versus those resulting in contact sensitization. One area of promising research directed toward this end is that in which the cytokine "profiles" resulting from exposure to irritants or contact sensitizers are being defined. That cytokine "profiles" parallel differences in type of toxic response (dermal irritation vs. contact sensitization) has been demonstrated in several of the research studies discussed above. Additional recent work by Dearman and Kimber (1999) demonstrates that differences in cytokine expression are also apparent in dermal versus respiratory sensitization responses. Hence, particular cytokine "profiles" may well represent measurable endpoints unique to the contact sensitization response that can be integrated effectively into a replacement assay.

References

Buehler, E. V. (1965). Delayed contact hypersensitivity in the guinea pig. *Arch. Dermatol. 91*, 171-177.

Coquette, A., Berna, N., Vandenbosch, A. et al. (1999). Differential expression and release of cytokines by an in vitro reconstructed human epidermis following exposure to skin irritant and sensitizing chemicals. *Toxicol. In Vitro 13(6)*, 867-877.

Cornacoff, J. B., House, R. V. and Dean, J. H. (1988). Comparison of a radioisotopic incorporation method and the mouse ear swelling test (MEST) for contact sensitivity to weak sensitizers. *Fund. Appl. Toxicol. 10*, 40-44.

Cumberbatch, M., Dearman, R. J. and Kimber, I. (1997). Langerhans cells require signals from both tumour necrosis factor-alpha and interleukin-1 beta for migration. *Immunol. 92(3)*, 388-395.

Cumberbatch, M., Dearman, R. J. and Kimber, I. (1999). Inhibition of dexamethasone of Langerhans cell migration: influence of epidermal cytokine signals. *Immunopharmacol. 41(3)*, 235-243.

Dearman, R. J., Hilton, J., Basketter, D. A. and Kimber, I. (1999). Cytokine endpoints for the local lymph node assay: consideration of interferon-gamma and interleukin 12. *J. Appl. Toxicol. 19(3)*, 149-155.

Dearman, R. J. and Kimber, I. (1999). Cytokine fingerprinting: characterization of chemical allergens. *Toxicol. Methods 19(1)*, 56-63.

De Silva, O. et al. (1996). Alternatives methods for skin sensitisation testing. The report and recommendations of ECVAM workshop 19. *ATLA 24(5)*, 683-705.

Dunn, B. J., Rusch, G. M., Siglin, J. C. and Blaszcak, D. L. (1990). Variability of a mouse ear swelling test (MEST) in predicting weak and moderate contact sensitization. *Fund. Appl. Toxicol. 15*, 242-248.

ECVAM (European Center for the Validation of Alternative Methods) (2000). Statement on the scientific validity of local lymph node assay for skin sensitisation testing.

Elmets, C. A. (1996). Progress in the development of an in vitro assay for contact allergens: Results of the Avon Program Project Experience. *In Vitro Toxicol. 9 (2)*, 223-235.

Enk, A. H. and Katz, S. I. (1992). Early molecular events in the induction phase of contact sensitivity. *Proc. Nat. Acad. Sci. (USA) 89*, 1398-1402.

Enk, A. H., Angeloni, V. L., Udey, M. C. and Katz, S. I. (1993). An essential role for Langerhans cell-derived IL-1ß in the intiation of primary immune responses in skin. *J. Immun. 150*, 3698-2704.

EU (European Union) (1992). Annex to Commission Directive 92/69/EEC of 31 July 1992 adapting to technical progress for the seventeenth time Council Directive 67/548/EEC on the approximation of laws, regulations, and administrative provisions relating to the classification, packaging, and labelling of dangerous substances. *Official Journal of the European Communities L383A*, 131-136.

Gad, S. C., Dunn, B. J., Dobbs, D. W et al. (1986). Development and validation of an alternative dermal sensitization test: the mouse ear swelling test. *Toxicol. Appl. Pharmacol. 84*, 93-114.

Hariya, T., Hatao, M. and Ichikawa, H. (1999). Development of a non-radioactive endpoint in a modified local lymph node assay. *Food Chem. Toxicol. 37(1)*, 87-93.

Hayes, B. B., Gerber, P. C., Griffey, S. S. and Meade, B. J. (1998). Contact hypersensitivity to dicyclohexylcarbodiimide and diisopropylcarbodiimide in female B6C3F1 mice. *Drug Chem. Toxicol. 21(2)*, 195-206.

Hayes, B. B. and Meade, B. J. (1999). Contact sensitivity to selected acrylate compounds in B6C3F1 mice: relative potency, cross reactivity, and comparison of test methods. *Drug Chem. Toxicol. 22 (3)*, 491-506.

Hoppe, U., Degwert, J. and Steckel, F. (1995). Use of CD1a - dendritic cells and keratinocytes to characterize cellular reactions involved in allergic contact dermatitis. *J. Cell. Biochem. 21 (Suppl. A)*, 11.

Howell, M. D., Manetz, T. Z. and Meade, B. J. (2000). Comparison of murine assays for the identification of chemical sensitizers. *Toxicology Methods 10(1)*, 1-15.

ICCVAM. (Interagency Coordinating Committee for the Validation of Alternative Methods) and NICEATM. (National Interagency Committee for the Evaluation of Alternative Methods) (1999). The Murine Local Lymph Node Assay. A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds. February. NIH Publication No. 99-4494.

Kimber, I., Hilton, J. and Botham, P. A. (1990). Identification of contact aller-

gens using the murine local lymph node assay. Comparisons with the Buehler occluded patch test in guinea pigs. *J. Appl. Toxicol. 10*, 173-180.

Kimber, I., Hilton, J., Botham, P. A. et al. (1991). The murine local lymph node assay. Results of an inter-laboratory trial. *Toxicol. Lett. 55*, 203-213.

Kimber, I. and Basketter, D. A. (1992). The murine local lymph node assay: A commentary on collaborative studies and new directions. *Fd. Chem. Toxicol. 30*, 165-169.

Kimber, I, Hilton, J., Dearman, R. J. et al. (1995). An international evaluation of the murine local lymph node assay and comparison of modified procedures. *Toxicology 103*, 63-73.

Kobayashi, Y., Staquet, M. J., Dezutter-Dambuyant, C. and Schmitt, D. (1994). Development of motility of Langerhans cells through extracellular matrix by in vitro hapten contact. *European J. of Immunol. 24*, 2254-2257.

Krasteva, M., Peguet-Navarro, J., Moulon, C. et al. (1996). In vitro primary sensitization of hapten-specific T cells by cultured human epidermal Langerhans cells – a screening predictive assay for contact sensitizers. *Clin. Exp. Allergy 26(5)*, 563-570.

Lenz, A., Heine, M., Schuler, G. and Romani, N. (1993). Human and murine dendritic cells. Isolation by means of a novel method and phenotypical and functional characterizations. *J. Clin. Invest. 95*, 258-2596.

Le Poole, I. C., Das, P. K., Krieg, S. R. et al. (1991). Organotypic culture of human skin for studying wound healing. *Wounds 3*, 102-110.

Magnusson, B. and Kligman, A. M. (1970). *Allergic Contact Dermatitis in the Guinea Pig.* Springfield, Illinois: Charles C. Thomas.

Messadi, D. V., Pober, J. S. and Murphy, G. F. (1988). Effects of recombinant gamma-interferon on HLA-DR and DQ in short term skin organ culture. *Lab. Invest. 58*, 61-67.

Moulon, C., Peguet-Navarro, J., Courtellemont, P. et al. (1993). In vitro primary sensitization and restimulation of hapten-specific T cells by fresh and cultured human epidermal Langerhans'cells. *Immunol. 80(3)*, 373-379.

Muller-Decker, K., Furstenberger, G. and Marks, F. (1994). Keratinocyte-derived proinflammatory key mediators and cell viability as in vitro parameters of irritancy: a possible alternative to the Draize skin irritation test. *Toxicol. Appl. Pharmacol. 127(1)*, 99-108.

Ohta, K., Yamanaka, S. and Takaesu, Y. (1999). The evaluation of mercury sensitization by mouse ear swelling test (MEST). *Shikwa Gakuho 99 (10)*, 821-832.

OECD (Organization for Economic Co-operation and Development) (1993). OECD Guideline for Testing of Chemicals, No. 406. Skin Sensitisation, 9 pp. Paris, OECD.

Pichowski, J. S., Cumberbath, M., Dearman, R. J. et al. (2000). Investigation of induced changes in interleukin 1beta mRNA expression by cultured human dendritic cells as an in vitro approach to skin sensitization testing. *Toxicol. Vitr. 14(4)*, 351-360.

Reutter, K., Jaeger, D., Degwert, J. and Hoppe, U. (1997). In vitro model for contact sensitization: II. Induction of IL-1beta mRNA in human blood-derived dendritic cells by contact sensitizers. *Toxicol. In Vitro 11(5)*, 619-626.

Romani, N., Gruner, S., Brang, D. et al. (1994). Proliferating dendritic cell progenitors in human blood. *J. Exp. Medicine 180*, 83-93.

Rougier, N., Redziniak, G., Mougin, D. et al. (2000). In vitro evaluation of the sensitization potential of weak contact allergens using Langerhans-like dendritic cells and autologous T cells. *Toxicol. 145(1)*, 73-82.

Sailstad, D. M., Tepper, J. S., Doerfler, D. L. and Selgrade, M. K. (1993). Evaluation of several variations of the mouse ear swelling test (MEST) for detection of weak and moderate contact sensitizers. *Toxicol. Methods. 3*, 169-182.

Tammi, R. and Maibach, H. I. (1987). Skin organ culture: why? *Int. J. Dermatol. 26*, 150-160.

Thorne, P. S., Hawk, C., Kaliszewski, S. D. and Guiney, P. D. (1994). The non-invasive mouse ear swelling assay. I. Refinements for detecting weak sensitizers. *Fund. Appl. Toxicol. 17*, 790-806.

USEPA (United States Environmental Protection Agency) (1998). Health Effects Test Guidelines. OPPTS 870.2600. Skin Sensitization. August.

5 Alternatives to developmental/reproductive toxicity testing in animals

5.1 Background

Several years ago, Gray and Kimber (1989) reported on the value and limitations of several *in vitro* techniques for reproductive and developmental toxicity. The researchers represented in the report emphasized the overall difficulty associated with developing replacement assays for very complex reproductive and developmental events in the whole organism. Other researchers since that time have echoed these sentiments (for example, Mirkles, 1996). However, increased emphasis on the use of *in vitro* testing by regulators and industry alike, has spurred attempts to validate assays of reproductive and/or developmental effects that do not use animals.

Any *in vitro* assay that could replace current reproductive testing procedures using animals would substantially reduce the number of animals used for biomedical testing. For example, current testing guidelines from the United States Environmental Protection Agency (USEPA) for reproductive/developmental toxicity screening tests as well as prenatal developmental toxicity and reproduction and fertility effects studies recommend using ten or twenty animals (at minimum) per

sex per dose level. Three dose levels are required as well as a control group in most cases (USEPA, 1998a, 1998b, 2000a, 2000b). Hence, 80 to 160 animals per assay are required at minimum.

Clearly, alternative assays that do not use animals are needed. Four promising candidates, the micromass (MM) assay, the whole embryo culture (WEC) assay, the embryonic stem cell test (EST), and the frog embryo teratogenesis assay – *Xenopus* (FETAX) assay, are discussed in-depth below.

5.2 *In vitro* alternatives

5.2.1 Micromass (MM) assay

The MM assay may use cultures of limb bud cells or neuronal cells or both. Faustman (1994) described these cultures as follows. "Limb or cephalic tissues (usually mesencephalic regions) are isolated from early to mid-organogenesis embryos and single cell suspensions are prepared using a combination of mechanical dissociation and enzymatic digestion. Cells are plated at high density and undergo differentiation into chondrocytes or neurons without additional stimulation."

In one of the first investigations of a mammalian MM system, Flint (1983) described a micromass culture method for rat embryonic neural cells noting that "the formation of neurons, the binding of neural cell projections (axons) into nerve-like structures, the synthesis of neuronal cell antigen and (3H)GABA uptake studies confirm that the micromass culture system may be a very good model of *in vivo* neuronal differentiation".

Tab. 6: Concentrations of nonteratogens and teratogens inhibiting differentiation in CNS and LB cultures by 50% of the control value (IC50) and maximum concentration having no inhibitory effect (NEC) (from Flint and Orton, 1984)

	Without S9				With S9			
	CNS		LB		CNS		LB	
Compound	**IC50**	**NEC**	**IC50**	**NEC**	**IC50**	**NEC**	**IC50**	**NEC**
Nonteratogens								
Dimenhydrinate	7.0[a]	1.0	180.0	100.0	60.0	10.0	-[b]	-
Glutehimide	300.0	200.0	160.0	100.0	-	-	-	-
Teratogens								
Aldrin	25.4	1.0	14.6	1.0	43.0	10.0	57.0	1.0
Aspirin	190.0	100.0	190.0	100.0	-	-	-	-
8-Azaguanine	0.8	<0.1	0.8	<0.1	300.0	100.0	280.0	100.0
Azathioprene	162.0	100.0	9.0	<0.1	230.0	100.0	-	-
Cadmium chloride	2.5	1.0	3.0	1.0	27.0	10.0	55.0	10.0
Captan	0.4	<0.1	3.8	1.0	20.0	10.0	40.0	10.0
Chloramphenicol	230.0	100.0	160.0	60.0	-	-	-	-
Colchicine	<0.1	<0.1	<0.1	<0.1	0.3	0.1	0.3	0.1
Cycloheximide	<0.1	<0.1	<0.1	<0.1	0.4	0.1	0.9	0.1
Cyclophosphamide	325.0	100.0	-	-	<0.1	<0.1	<0.1	<0.1
Cytosine arabinoside	<0.1	<0.1	<0.1	<0.1	0.8	0.1	3.0	0.1
Diazepam	150.0	100.0	150.0	100.0	-	-	-	-
Diethylstilbestrol	2.8	1.0	1.8	0.1	-	-	-	-
Diphenylhydantoin	64.5	25.0	78.5	50.0	250.0	100.0	-	-
L-Dopa	40.0	10.0	40.0	10.0	-	-	0.4	<0.1
Ethionine	-	-	-	-	-	-	<0.1	<0.1
EDTA	2.8	1.0	-	-	-	-	-	-
5-Fluorouracil	<0.1	<0.1	<0.1	<0.1	0.3	<0.1	0.2	<0.1
Hydrocortisone	200.0	40.0	100.0	30.0	-	-	-	-
Indomethacin	45.0	10.0	50.0	10.0	240.0	100.0	-	-
Mebendazole	<0.1	<0.1	<0.1	<0.1	4.5	1.0	1.6	0.1
Parbendazole	<0.1	<0.1	0.9	<0.1	-	-	-	-
Sulfisoxazole	160.0	100.0	210.0	10.0	-	-	-	-
Theophylline	-	-	140.0	100.0	-	-	-	-
ICI 124,206	<0.1	<0.1	<0.1	<0.1	3.5	1.0	3.5	1.0

Note: The following nonteratogens did not inhibit differentiation at any of the concentrations used: aminopyrine ascorbic acid, caffeine, l-carnosine, cyclamate, inosine, kanamycin sulfate, maltose, 6-methyluracil, nitrilotriacetic acid, quinine, saccarin, sulfanilamide, sulfisomidine, tolbutamide, ICI 89406 and ICI 123,737. Two teratogens: 2,4-Dichlorophenoxyacetic acid and thalidomide, similarly did not inhibit differentiation.

[a] µg/ml culture medium

[b] No inhibitory effect to 500 µg/ml or maximum soluble concentration (if less).

Shortly thereafter, Flint and Orton (1984) published results from a study in which 46 compounds (27 teratogens/19 nonteratogens) were investigated by *in vitro* assay using cultures of rat embryo midbrain and limb bud cells (Tab. 6).

A high specificity (89% prediction of nonteratogens) and sensitivity (93% prediction of teratogens) were noted in this study verifying a high predictability for this MM assay.

Subsequently, Renault and colleagues (1989) investigated the micromass limb bud cell culture as a model for studying teratogenic mechanisms. They monitored changes in cell differentiation and cell proliferation after treatment with teratogens, retinoic acid and 6-aminonicotinamide, and the non-teratogen, doxylamine. Retinoic acid inhibited cell differentiation whereas 6-aminonicotinamide inhibited cell proliferation. Doxylamine produced a non-specific cytotoxic effect. The assay was subsequently refined to a routine *in vitro* teratogenesis screening test in which cells were cultivated in microtiter plates and chondrogenesis and cell proliferation were simultaneously assessed after six days of culture. The test was validated with 51 reference compounds, 33 teratogens and 18 non-teratogens. None of the non-teratogens inhibited chondrogenesis or cell proliferation. Further ultrastructural characterization of normal and abnormal chondrogenesis in micromass rat embryo limb bud cell cultures was also performed by these researchers (Renault et al., 1995).

Characterization of neuronal micromass cultures has been performed by Whittaker et al. (1993). They concluded that cells taken from gestational day-12 embryos and cultured for 5 days *in vitro* comprise at least two morphologically distinct cell types: fibroblast-like cells and neurons. Antibodies to GQ ganglioside, g-aminobutyric acid (GABA), microtubule-associated protein 2 (MAP2), MAP5, neuron-specific enolase (NSE), neural cell adhesion molecules (N-CAM), and tau yielded preferential staining. Hence, these workers concluded that the pattern of neuronal differentiation in CNS micromass cultures exhibits certain similarities to that observed *in vivo*. In addition, they felt that certain markers identified in the study were of potential utility as biomarkers of chemically-induced developmental neurotoxicity.

Since the development of the MM assay, it has been used widely to investigate the developmental toxicity of numerous chemicals. Examples include, acrylamide, methylene bisacrylamide and haloperidol (Walum and Flint, 1993); chlorpyrifos (Cosenza and Bidanset, 1995); paraquat (Cosenza, et al., 1996); glufosinate ammonium (Watanabe and Iwase, 1996); acetylsalicylic acid, isoniazid, penicillin G, saccharine, vincristine sulphate, 6-aminonicotinamide, retinoic acid, and amaranth (Amacher et al., 1996); methylmercury (McMillan, et al., 1997); and cadmium (Haberstroh and Kapron, 1998). Recent evaluations of the MM system by an ECVAM expert review (Brown et al., 1995) indicated that these cultures "represent robust test systems for studying potential teratogens". They cited the fact that there is probably more experience with this system than with any other alternative method for reproductive toxicology studies and that its limitations are well understood.

Tab. 7: Toxicity of phenol, chlorophenols and selected derivates in the adult Hydra (from Mayura et al., 1991)

Chemical	Adult hydra toxicity whole-log MAC[a] (µg/ml)	Exposure time[b]
Phenol	1000	20
Anisole	1000	44
Phenyl acetate	100	20
Phenol. Na salt	1000	28
4-Chlorophenol	100	4
3-Chlorophenol	100	20
2-Chlorophenol	100	92
4-Chlorophenyl acetate	100	4
4-Chlorophenol, Na salt	100	4
2,3-Dichlorophenol	100	4
2,4-Dichlorophenol	10	92
2,5-Dichlorophenol	10	92
2,6-Dichlorophenol	100	20
3,4-Dichlorophenol	10	44
3,5-Dichlorophenol	10	20
3,5-Dichlorophenyl acetate	10	20
3,5-Dichlorophenol, Na salt	10	20
2,3,4-Trichlorophenol	1	44
2,3,5-Trichlorophenol	1	20
2,3,6-Trichlorophenol	10	68
2,4,5-Trichlorophenol	1	28
2,4,6-Trichlorophenol	10	28
3,4,5-Trichlorophenol	1	44
2,3,5-Trichlorophenyl acetate	1	28
2,3,5-Trichlorophenol, Na salt	1	28
2,3,4,5-Tetrachlorophenol	0.1	44
2,3,4,6-Tetrachlorophenol	1	44
2,3,5,6-Tetrachlorophenol	1	44
2,3,4,5-Tetrachloroanisole	10	92
2,3,5,6-Tetrachloroanisole	10	68
2,3,4,5-Tetrachlorophenyl acetate	0.1	44
2,3,4,5-Tetrachlorophenol, Na salt	0.1	44
Pentachlorphenol	0.1	44
Pentochloranisole	10	92
Pentachlorphenyl acetate	0.1	44
Pentachlorphenol, Na salt	0.1	44
Tetrachloro-1,2-benzoquinone	1	92
Tetrachloro-1,4-benzoquinone	10	20

[a] MAC (Minimum effective concentration)
[b] Time (hr) required to reach toxic endpoint

5.2.2 Whole embryo culture (WEC) assay

Mammalian embryo cultures are derived from mice, rats, or rabbits as summarized by Brown et al. (1995). Head-fold or early somite stage embryos are dissected free from maternal tissue, parietal yolk sac and Reichert's membrane, leaving the visceral yolk sac and ectoplacental cone intact. The conceptus is cultured in medium under defined gassing conditions for 24-48 hours. Media containing a high proportion of serum is usually used and test compounds are added to the cultures for defined periods of time. Metabolic activation systems may also be incorporated. A number of endpoints can be measured such as effects on development of the visceral yolk sac vascularisation and circulation; effects on haematopoiesis, embryonic growth and differentiation; and dysmorphogenic effects.

Several studies have investigated the WEC assay in relationship to other assays of developmental toxicity. Bechter and colleagues (1991) investigated the ability of rat WEC to differentiate between retinoids with differing teratogenic potency. They compared the results obtained to those from *in vivo* studies of mice and micromass cultures of rat embryonic limb bud cells and concluded that the WEC "was an excellent tool for the preliminary testing of retinoids".

The developmental toxicity of chlorinated phenols was evaluated by Mayura and co-workers (1991) utilizing *Hydra attenuata* and rat embryos in culture. The WEC findings closely paralleled those found using Hydra embryos (Tab. 7, 8, and 9) and the authors suggested that a combination of Hydra and WEC assays would facilitate rapid detection and ranking of hazardous chemicals associated with complex mixtures of chemical wastes.

Gregotti et al. (1992) evaluated the effects of styrene oxide in a limb bud micromass assay and a WEC system. Styrene oxide induced a dose-dependent increase in embryolethality and embryo

Tab. 8: Comparison of phenol and chlorinated phenols at equimolar concentration (0.6mM) on growth and development of day 10 rat embryos in cultures (from Mayura et al., 1991)

Treatment	Number of embryos exposed	Yolk-sac diameter* (mm)	Crown-rump length* (mm)	Number of somites* (AV)	Morphological score
Control	26	7.4 ± 0.13[a]	4.6 ± 0.09[a]	32.4 ± 0.33[a]	49.3 ± 0.52[a]
Solvent control (0.5% DMSO)	21	7.4 ± 0.13[a]	4.2 ± 0.12[a]	32.5 ± 0.36[a]	48.3 ± 0.49[a]
Phenol	10	7.0 ± 0.20[a]	4.6 ± 0.21[a]	31.5 ± 0.70[a]	48.9 ± 0.18[a]
4-Chlorophenol	12	7.3 ± 0.11[a]	3.9 ± 0.17[b]	29.5 ± 0.65[b]	45.6 ± 0.93[b]
3,5-Dichlorophenol	14	5.4 ± 0.12[b]	2.4 ± 0.09	20.0 ± 0.65[c]	17.8 ± 0.75[c]
2,3,5-Trichlorophenol	13	4.8 ± 0.43[b]	2.9 ± 0.14[c]	19.6 ± 0.60[c]	24.8 ± 1.12[c]
2,3,4,5-Tetrachlorophenol	7	3.7 ± 0.90[c]	2.3 ± 0.06[d]	very few	5.0 ± 2.00[f]
Pentachlorophenol	14	5.3 ± 0.36[b]	3.1 ± 0.10[c]	20.9 ± 0.46[c]	21.5 ± 0.27[d]

Note: Means with the same letter are not significantly different ($p \leq 0.05$)
*****Values are means ± standard error

Tab. 9: Comparison of phenol and chlorinated phenols at equimolar concentration (0.6mM) on growth and development of day 10 rat embryos in cultures (from Mayura et al., 1991)

Treatment	Number of embryos exposed	Protein$^{\Delta}$ µg/embryo	DNA$^{\Delta}$ µg/embryo	Number malformed
Control	26	298.31 ± 17.2[a]	65.2 ± 3.2[a]	0
Solvent control (0.5% DMSO)	21	287.55 ± 23.9[a]	62.0 ± 3.5[a]	0
Phenol	10	294.00 ± 18.5[a]	67.2 ± 4.1[a]	0
4-Chlorophenol	12	195.00 ± 18.7[b]	47.6 ± 3.8[b]	0
3,5-Dichlorophenol	14	72.14 ± 7.0[c,d]	20.6 ± 1.7[c,d]	100%
2,3,5-Trichlorophenol	13	83.50 ± 6.9[c]	21.8 ± 1.5[c,d]	*
2,3,4,5-Tetrachlorophenol	7	24.50 ± 3.6[d]	14.7 ± 2.5[d]	*
Pentachlorophenol	14	127.50 ± 11.1[c]	26.5 ± 1.1[c]	*

Note: Means with the same letter are not significantly different ($p \leq 0.05$)
$^{\Delta}$ Values are means ± standard error
***** Inhibition of growth and development

malformation at concentrations that were comparable to effective concentrations in the micromass culture system. These researchers felt that these findings indicated a "quantitative comparability between the two *in vitro* systems analyzed".

The WEC assay has also been used effectively to differentiate the stereoselective teratogenicity of enantiomers of the valproic acid analogue, 2-n-propyl-4-pentynoic acid (4-yn-VPA). Andrews et al. (1995) studied the effects of the R(+) and S(-) enantiomers of 4-yn-VPA on cultures of mouse embryo. The teratogenic effects of the S(-) enantiomer were clearly visible in this assay, whereas, the R(+) enantiomer produced no embryotoxic nor dysmorphogenic effects at any concentration (Tab. 10 and 11).

This study represented the first illustration of the stereoselectivity of a developmental toxicant in an *in vitro* test system, mirroring results obtained in an *in vivo* assay.

ECVAM has indicated that this *in vitro* system is a valuable tool considering its extensive use in many laboratories and its ability to detect dysmorphogenesis in embryos.

Moreover, they noted that it enables not only the detection of adverse developmental effects from chemical exposure, but has also been shown to discriminate between the potencies of structurally related compounds. They caution, however, that the system is limited in that it is complex, covers only a part of organogenesis (a feature shared by other *in vitro* assays of developmental/reproductive effects), and requires a fairly high level of technical sophistication to perform (Brown et al., 1995).

5.2.3 Embryonic stem cell (EST) test

One of the major drawbacks of micromass culture systems is that they require the primary culture of cells isolated from embryos. Although avian embryos can be used in place of mammalian embryos, the isolation of fresh cells is still a limiting factor in the development of these systems. Researchers have noted, however, that embryonic stem (ES) cell lines can be established from mammalian blastocysts. Essentially, embryonic stem cells can be maintained in an undifferentiated state in the presence of feeder layers and/or purified leukemia inhibitory factor (LIF). Upon removal of the LIF, ES cells differentiate into a variety of cell types (Brown et al., 1995). Optimization of culture conditions for ES cells has been studied by several laboratories (Laschinski et al., 1989; Mummery et al., 1993) and attempts at validation of the assay as predictive of teratogenicity have been undertaken.

In one such study, Newall and Beedles (1994) examined the effects of 25 compounds on mouse ES cells and compared them to results obtained with rat primary embryonic cells in a micromass test protocol. The stem-cell test predicted the teratogenicity of these compounds with a similar sensitivity and specificity to the micromass test. The authors noted that an added advantage of the stem-cell test was that "the test used a propagated cell line; there was no use of animals". This particular method utilized colony formation as a measurable endpoint of adverse effect. Subsequently, as discussed below, other endpoints have been shown to be more useful as predictive tools.

Tab. 10: The effects of S-4-yn-VPA on embryonic development in whole-embryo culture[a] (from Andrews et al., 1995)

Concentration (mmol/liter)	Yolk-sac diameter (mm)	Crown-rump length (mm)	Head length (mm)	Somite number	Developmental score	Dead (%)	Abnormal (%)	Ending pH
0	3.4 ± 0.08	2.8 ± 0.07	1.35 ± 0.05	21 ± 0.5	41 ± 1.3	7	15	7.09 ± 0.04
0.075	3.4 ± 0.08	2.8 ± 0.06	1.34 ± 0.04	20 ± 0.5	41 ± 1.1	0	38	7.13 ± 0.05
0.15	3.3 ± 0.07	2.7 ± 0.07	1.31 ± 0.04	21 ± 0.5	38 ± 1.1	0	50*	7.13 ± 0.05
0.3	3.2 ± 0.08	2.6 ± 0.06	1.24 ± 0.04	19 ± 0.5	38 ± 1.1	0	56*	7.13 ± 0.05
0.6	3.2 ± 0.08	2.4 ± 0.06*	1.06 ± 0.04*	18 ± 0.5*	31 ± 1.1*	0	94*	7.19 ± 0.05
1.2	3.1 ± 0.13	2.3 ± 0.15*	0.93 ± 0.08*	16 ± 1.3*	25 ± 2.0*	69*	100*	7.28 ± 0.05

[a] Values are means ± SE for 13 - 16 embryos.
* $p \leq 0.05$ compared to controls.

Tab. 11: The effects of R-4-yn VPA on embryonic development in whole-embryo culture[a] (from Andrews et al., 1995)

Concentration (mmol/liter)	Yolk-sac diameter (mm)	Crown-rump length (mm)	Head length (mm)	Somite number	Developmental score	Dead (%)	Abnormal (%)	Ending pH
0	3.3 ± 0.08	2.7 ± 0.07	1.4 ± 0.05	21 ± 0.5	41 ± 1.1	6	19	7.05 ± 0.05
0.075	3.3 ± 0.09	2.7 ± 0.07	1.3 ± 0.04	20 ± 0.5	40 ± 1.3	14	21	7.05 ± 0.05
0.15	3.3 ± 0.07	2.7 ± 0.07	1.3 ± 0.04	19 ± 0.5	39 ± 1.1	0	44	7.07 ± 0.05
0.3	3.3 ± 0.07	2.7 ± 0.06	1.3 ± 0.04	20 ± 0.5	41 ± 1.1	0	25	7.07 ± 0.05
0.6	3.3 ± 0.08	2.6 ± 0.07	1.4 ± 0.04	20 ± 0.5	40 ± 1.1	6	31	7.12 ± 0.05
1.2	3.2 ± 0.07	2.7 ± 0.07	1.2 ± 0.04	19 ± 0.5	39 ± 1.1	0	25	7.07 ± 0.05

[a] Values are means ± SE for 14-16 embryos.
* $p \leq 0.05$ compared to controls.

Utilizing a somewhat different protocol, Spielmann and colleagues (1997) evaluated an embryonic stem cell test composed of two permanent mouse cell lines, 3T3 fibroblasts and embryonic stem cells (D3). This evaluation investigated the effectiveness of 12 different endpoints and endpoint ratios in predicting the embryotoxic potential of 16 test compounds. Three endpoints (inhibition of differentiation of ES cells into cardiac myoblasts (identified by contraction) and cytotoxicity of ES cells and 3T3 cells (MTT-test)) were found to show a better correlation to the embryotoxic properties of the test chemicals than the other endpoints. Utilizing these endpoints, 50% inhibition concentrations for differentiation (ID_{50}) and cytotoxicity (IC_{50}D3 and IC_{50}3T3) were calculated from concentration-response curves. When coupled with linear discriminant analysis, these data formed the basis of a classification scheme in which test chemicals were assigned to three classes: embryotoxic, moderately embryotoxic and strongly embryotoxic. Using this model, all sixteen test chemicals were correctly assigned to their *in vivo* classes of embryotoxicity.

Subsequently, an interlaboratory comparison of results obtained using the stem cell test described above was performed (Spielmann et al., 1998; Scholz et al., 1999a). Three test chemicals representing three classes of embryotoxicity (non-, weak/moderate, and strong) were evaluated by three different laboratories as part of the protocol transfer phase (II) of prevalidation experiments. In the final phase (III) of prevalidation (protocol performance), nine test chemicals and a positive control of known embryotoxic potential *in vivo* and similar potential in animals and in humans were selected and tested by three laboratories. Because the original prediction model calculated for the EST was unsatisfactory, an improved iPM was utilized to analyze the results obtained from Phases I and II. The application of the iPM to phase III data from two of the participating laboratories provided correct predictivities of 93 and 94%, respectively. Application of the iPM to all data from the three laboratories involved in Phase II of the study provided a 100% correct prediction. Although there was some overlap in the chemicals tested in the different phases of the study, the iPM proved to be more appropriate than the original PM for data sets from different laboratories. Scholz et al. (1999b) discusses in-depth the improved prediction model (iPM). Unlike the original PM, the improved PM incorporated as one variable the relative

Tab. 12: Results of the interlaboratory study for Hydroxyurea (from Bantle et al., 1994)

Laboratory	Replicate (n)	LC_{50} (mg ml^{-1})	Mean LC_{50} (mg ml^{-1})	LC_{50} CV[a] (%)	EC_{50} (mg ml^{-1})	Mean EC_{50} EC_{50} (mg ml^{-1})	EC_{50} CV(%)	Mean TI[b]	TI CV(%)	MCIG[c] (mg ml^{-1})	Mean MCIG (mg ml^{-1})	MCIG CV(%)
1	1	0.45			0.16					0.25		
	2	0.49	0.45	7.3	0.15	0.16	7.8	2.8	14.5	0.15	0.18	25.7
	3	0.41			0.18					0.15		
2	1	0.71			0.17					0.10		
	2	0.32	0.51	31.2	0.10	0.15	23.7	3.4	17.3	0.15	0.15	27.2
	3	0.50			0.18					0.20		
3	1	0.45			0.21					0.40		
	2	2.19	1.11	69.5	0.41	0.23	58.4	4.8	54.1	0.50	0.53	23.4
	3	0.68			0.08					0.70		
4	1	0.80			0.14					0.10		
	2	d	1.08	25.9	0.22	0.19	18.7	5.7	7.0	0.15	0.17	37.4
	3	1.36			0.21					0.25		
5	1	0.23			0.07					0.20		
	2	0.12	0.23	40.8	0.12	0.11	26.8	2.1	45.6	0.30	0.30	27.2
	3	0.35			0.14					0.40		
6	1	1.82			0.28					0.20		
	2	1.24	1.56	15.5	0.33	0.26	24.0	6.0	36.1	0.50	0.28	54.6
	3	1.63			0.18					0.15		
7	1	0.55			0.15					0.15		
	2	0.40	0.44	18.6	0.12	0.13	10.9	3.4	8.5	0.10	0.12	20.2
	3	0.36			0.12					0.10		

[a] Coefficient of variation
[b] Mean teratogenic index (TI) calculated by dividing the mean LC_{50} by the mean EC_{50} for each laboratory
[c] Minimum concentration to inhibit growth
[d] An LC_{50} could not be calculated because there was insufficient mortality in the concentration range tested

distance between IC_{50}3T3 and ID_{50}, instead of the ratio ID_{50}/IC_{50}D3 that was used previously.

Brown et al. (1995) noted that "ES cells offer several new approaches with respect to screening for embryotoxicity *in vitro*, enabling the use of differentiating permanent embryonic cells". They also suggested that continued development of this assay should be supported. The endpoints used in the EST to date, those of cytotoxicity and effects on differentiation, have correlated reasonably well with results obtained with *in vivo* teratogens. Many researchers feel, however, that there are relevant endpoints yet to be discovered with this assay perhaps in conjunction with the use of transgenic animals, novel immunologic staining methods and/or studies of changes in gene expression in ES cells.

5.2.4 Frog embryo teratogenesis assay (*Xenopus*) (FETAX)

Unlike the three assays discussed above, FETAX evaluates the effect of toxicants on the development of embryos from a non-mammalian species, the South African clawed frog (*Xenopus laevis*). Originally developed by Dumont (1983), the assay has undergone a number of interlaboratory validation studies (Bantle et al., 1994a, 1994b, 1996, 1999) (Tab. 12). Moreover, in 1991, the American Society for Testing and Materials (ASTM) developed a test guideline for FETAX, which was subsequently revised and republished in 1998 (ASTM, 1991, 1998). Recent evaluations of FETAX data have focused on the utility of the assay as a screening protocol for developmental/reproductive toxicity testing in mammals as well as for ecotoxicological studies of potential reproductive hazard (NICEATM, 2000).

The rationale behind using FETAX as a monitor for teratogenic events in mammals is that embryonic development during the first 96 hrs in *Xenopus* parallels many of the major processes of human organogenesis. FETAX measures mortality, malformations and growth inhibition in a 96-hour *in vitro* whole-embryo test. Functional deficits resulting from exposure to a teratogenic substance are not monitored in FETAX.

However, metabolic activation systems can be incorporated into FETAX allowing for evaluation of substances that may require metabolism in order to produce teratogenicity. The concentration estimated to induce lethality in 50% of the exposed embryos (LC_{50}) and the concentration estimated to induce malformations in 50% of the exposed embryos (EC_{50}) are calculated based on mortality and malformation findings. The teratogenic index (TI) (LC_{50} divided by the EC_{50}) is also calculated. Growth is estimated by measuring the head to tail length of the embryos. The minimum concentration to inhibit growth (MCIG) is determined by comparing the mean 96-hour head to tail length of the treated embryos at each treatment concentration to that of the control embryos. Any one of these three criterion (TI, growth inhibition, or severity of induced malformations) is used to identify a test material as teratogenic.

5.3 Validation and regulatory activities

The Organization for Economic Cooperation and Development (OECD) guidelines for developmental toxicity do not explicitly restrict developmental toxicity testing to mammals (OECD, 1981, 1983a, 1983b, 1995, 1996). However, most current regulatory guidelines for developmental/reproductive toxicity testing include tests that use animals, particularly mammals, when considering potentially adverse effects on humans. As noted above, the complexity of the developmental and reproductive systems in humans and other animals has hindered the search for an effective alternative assay or battery of assays. This hindrance, not surprisingly, is reflected in the current lack of strong regulatory initiatives supporting replacement alternatives to developmental/reproductive toxicity testing in animals. However, use of alternative assays as screening tests for reproductive toxicants seems to be gaining a foothold as validation efforts provide additional evidence that several of these assays are indeed suitable predictors of *in vivo* reproductive toxicity.

ECVAM (European Center for the Validation of Alternative Methods) is continuing work with a full validation of the MM and WEC assays as well as the EST (Prieto, 2000). Moreover, Genschow et al. (2000) cite encouraging results regarding development and use of prediction models for these assays. The PMs generated for the MM and WEC assays, provided 81% (MM) and 72% (WEC) correct classifications. The improved PM (iPM) for the EST resulted in 79% correct classifications when compared to embryotoxic potential defined by *in vivo* testing. Although, the results given for the MM and WEC assays may be modified downward when the entire contingent of 20 chemicals is tested, the percentages obtained thus far indicate a reasonable degree of correlation between these three *in vitro* assays and their *in vivo* counterparts.

An important, although often overlooked, aspect of validation experiments, is that they often serve to refine the protocols undergoing validation. Examples of this are found in results from work with both the MM and WEC assays as cited by Genschow. Analysis of MM data indicated that the midbrain cells in the MM assay did not provide reproducible data. Hence, prevalidation of this assay only includes data obtained from limb bud cells. Furthermore, because the concentration response curves of the differentiation of limb bud cells monitored by AB staining and cytotoxicity assessed by neutral red uptake were almost identical, only differentiation of limb bud cells was considered because it showed a higher predictive power than cytotoxicity. Moreover, although a wide spectrum of results is available for evaluation in the WEC assay (i.e. 16 parameters for general morphological development and 30 types of malformations), only morphological development (total morphological score) and the percentage of malformed embryos provided sufficient information for predicting embryotoxic potential in a PM. This type of refinement (toward more relevant endpoints for prediction modeling) tends to simplify assay procedures. Ultimately, this simplification should enhance the effectiveness of *in vitro* assays, whether they monitor developmental and/or reproductive toxicity or other types of toxicity.

The USEPA has also recently requested that the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) evaluate the validation status of FETAX. In response to this request, ICCVAM in conjunction with NICEATM (National Toxicology Program Interagency Center for the Evaluation of Alternative Testing Methods) has finished a critical evaluation of FETAX (NICEATM, 2000). In this evaluation, NICEATM assessed the status of FETAX as a screening assay for detection of potential human teratogens. FETAX has been proposed as a screening test that, if positive, would indicate a potential human hazard, and, in the absence of other data, would be considered a presumptive teratogen/developmental toxicant. A negative FETAX response would not necessarily indicate the absence of a hazard, and negative responses would be followed by definitive *in vivo* mammalian testing. NICEATM compiled information from 276 studies involving 137 substances as it pertains to human development hazard assessment.

FETAX performance characteristics (i.e. accuracy, sensitivity, specificity, positive predictivity, negative predictivity, false positive rate, and false negative rate) against rat, mice, and/or rabbit teratogenicity test results or human teratogenicity study results were determined. The decision criteria used in determining the performance characteristics of FETAX included single decision criteria (TI > 1.5; TI > 3.0; MCIG/LC_{50} < 0.30) and multiple decision criteria (TI > 1.5 plus MCIG/LC_{50} < 0.30; TI > 3.0 plus MCIG/LC_{50} < 0.30). When a multiple decision criterion was used, test substances were classified as positive when both the TI values were greater than the decision point (1.5 or 3.0) and the MCIG/LC_{50} ratio was less than 0.3; equivocal when one, but not both, criterion were positive; and negative when neither criterion was positive.

The values obtained from this analysis suggested that the use of FETAX as a screen, based on current decision criteria, may pose some problems, although the assay is considered robust in many respects. Three approaches were advanced for modifying decision criteria to increase the ability of FETAX to correctly identify developmental toxicants. The first approach would be to evaluate the EC_{50} based on characteristic malformations (i.e., those increasing in incidence and severity with increasing test substance concentration) only. (Note: An atlas addressing *Xenopus* malformations for use with FETAX has recently been published (Bantle et al., 1998)). The second approach is to calculate a point estimate for the dose that inhibits growth by 50% rather than using an MCIG. The third approach would be to use 95% confidence intervals for statistically identifying TI values (and other point estimates) that are significantly different from the decision criteria value. This report also found excessive variability in the LC_{50}, EC_{50}, TI, and especially the MCIG within and across laboratories participating in the validation studies conducted with FETAX thus far. No immediate reason for this variability is apparent, although the lack of classification of malformations as to type and severity seems to be a contributing factor. On a more positive note, NICEATM's evaluation indicated that the projected cost (<$25,000) and study duration (less than two months), for a Good Laboratory Practice (GLP) compliant FETAX study, with and without metabolic activation were reasonable in comparison to the cost of a complete rat Prenatal Developmental Toxicity Study (approximately $120,000).

Overall, validation of several *in vitro* assays of developmental/reproductive toxicity seems to be reaching a critical culmination.

5.4 Summary, conclusions, and future work

In summary, validation efforts are progressing well for four *in vitro* assays of developmental/reproductive toxicity. Results from evaluations of the MM and WEC assays as well as the EST appear to be favorable; data from studies of FETAX suggest that further improvements in the assay would yield greater predictivity. The fact that the OECD does not restrict its reproductive testing guidelines to mammals, coupled with recent initiatives on the part of the USEPA seem to suggest that regulatory support of alternatives assays of development and reproductive toxicity is present, although, perhaps not as vigorous as it could be. This reticence may reflect both the complexity of the reproductive cycle and the need for validated *in vitro* assays, which are, as suggested by the discussion above, "on the horizon".

The next five years of research in this area should yield assays that are acceptable to both scientific and regulatory communities as screening tools for developmental and reproductive toxicity. This acceptance should promote much more widespread incorporation of screening batteries utilizing these assays into toxicity testing regimens for new compounds. Schwetz (1993) has pointed out, however, that an *in vivo* screen should perhaps be performed first with a small number of animals over a range of dose levels simply to see if the compound does indeed cause reproductive or developmental effects.

Over the next ten years, it may be possible to develop *in vitro* assays that represent much more selectively the various components of the reproductive cycle (i.e. the complexity of the system will be much more well-represented). This should help tremendously in identifying the exact locus of toxic action and in acceptance of *in vitro* assays as true replacement alternatives. Moreover, Brown et al. (1995) recommended that, "Methods using human semen for testing the effects of chemicals on mature sperm should be refined and validated".

Development of alternative assays of development and reproduction over the next twenty years may well incorporate "pattern of response" approaches such as those described in discussions of proteomics or toxicogenomics. Brown (1998) has already stated, "It is important that these new strategies (embryonic approaches) are not bedevilled by naive expectations, particularly in the early stages of their use. The V word (validation) should be locked away, in favor of "profiling", in which we ask 'Can this chemical affect this particular pathway?'" Moreover, he predicts that, "Once answers are available for many chemicals and pathways, patterns of response will be assembled, and these may allow the prediction of some types of developmental toxicity." These

thoughts are mirrored in the NICEATM report on FETAX (2000) that suggests that cDNA microarray technology would "greatly increase the utility of FETAX for identifying and prioritising developmental hazards". Hence, "profiling" of chemical pathways after reproductive toxicant exposure with cDNA microarray technology may well be the next definitive step in development of alternatives to the use of animals in developmental/reproductive toxicity testing.

References

Amacher, D., Stadler, J., Schomaker, S. and Verseil, C. (1996). The comparative testing of eight coded chemicals in the rat limb bud micromass and rat embryo culture systems. *Alternatives Lab. Animal 24 (6)*, 945-954.

Andrews, J. E., Ebron-McCoy, M., Bojic, U. et al. (1995). Validation of an in vitro teratology system using chiral substances: stereoselective teratogenicity of 4-yn-valproic acid in cultured mouse embryos. *Toxicol. Appl. Pharmacol. 132 (2)*, 310-316.

ASTM (American Society for Testing and Materials) (1991). Standard Guide for Conducting the Frog Embryo Teratogenesis Assay – Xenopus (FETAX). ASTM E1439–91. In *Annual Book of ASTM Standards*, Philadelphia, PA.

ASTM (1998). Standard Guide for Conducting the Frog Embryo Teratogenesis Assay – Xenopus (FETAX). ASTM E1439-98. In *Annual Book of ASTM Standards*, Philadelphia, PA.

Bantle, J. A., Burton, D. T., Dawson, D. A. et al. (1994a). Initial interlaboratory validation study of FETAX: Phase I testing. *J. Appl. Toxicol. 14*, 213-223.

Bantle, J. A., Burton, D. T., Dawson, D. A. et al. (1994b). FETAX interlaboratory validation study: Phase II testing. *Environ. Toxicol. Chem. 13*, 1629-1637.

Bantle, J. A., Finch, R. A., Burton, D. T. et al. (1996). FETAX interlaboratory validation study. Phase III, part 1 testing. *J. Appl. Toxicol. 16*, 517-528.

Bantle, J. A., Dumont, J. N., Finch, R. A. et al. (1998). *Atlas of Abnormalities: A Guide for the Performance of FETAX*, 2nd ed. Stillwater, OK: Oklahoma State University Press.

Bantle, J. A., Finch, R. A., Fort, D. J. et al. (1999). Phase III interlaboratory study of FETAX, part 3 – FETAX validation using 12 compounds with and without an exogenous metabolic activation system. *J. Appl. Toxicol. 19(6)*, 447-472.

Bechter, R., Terlouw, G. D., Tsuchiya, M. et al. (1991). Teratogenicity of arotinoids (retinoids): comparison of the whole embryo culture system with the in vivo mouse model and the limb bud cell culture. *Teratology 44(3)*, 29A.

Brown, N., Spielmann, H., Bechter, R. et al. (1995). Screening chemicals for reproductive toxicity: the current alternatives. The report and recommendations of ECVAM workshop 12. *ATLA 23*, 868-882.

Brown, N. A. (1998). Developmental toxicology in vitro: current status and imminent advances. *Hum. Exp. Toxicol. 17(9)*, 483.

Cosenza, M. and Bidanset, J. (1995). Effects of chlorpyrifos on neuronal development in rat embryo midbrain micromass cultures. *Vet. Hum. Toxicol. 37(2)*, 118-121.

Cosenza, M., Bidanset, J. and Johnson, D. E. (1996). In vitro/in vivo correlation of paraquat induced developmental neurotoxicity. *Toxicologist 30 (I Part 2)*, 244.

Dumont, J. N., Schultz, T. W., Buchanan, M. W. and Kao, G. L. (1983). III. Frog Embryo Teratogenesis Assay: Xenopus (FETAX) – A short-term assay applicable to complex environmental mixtures. In M. Waters, S. Sandhu, J. Lewtas, L. Claxton, N. Chernoff, and S. Nesnows (eds.), *Short-term Bioassays in the Analysis of Complex Environmental Mixtures* (393-405). New York: Plenum.

Faustmann, E. M. (1994). Mammalian micromass embryo cell cultures: applications in developmental toxicity evaluation and mechanistic investigations. *In Vitro Toxicol. 7(2)*, 158.

Flint, O. P. (1983). A micromass culture method for rat embryonic neural cells. *J. Cell Sci. 61*, 247-262.

Flint, O. P. and Orton, T. C. (1984). An in vitro assay for teratogens with cultures of rat embryo midbrain and limb bud cells. *Toxicol. Appl. Pharmacol. 76*, 383-395.

Genschow, E., Scholz, G., Brown, N. et al. (2000). Development of prediction models for three in vitro embryotoxicity tests in an ECVAM validation study. *In Vitro & Molecular Toxicology 13(1)*, 51-65.

Gray, T. and Kimber, I. (1989). Reproductive and developmental toxicology: the value and limitations of in vitro techniques. *Toxic. In Vitro 3(3)*, 241-242.

Gregotti, C., Manzo, L., Costa, L. G. and Faustman, E. M. (1992). Developmental toxicity of styrene oxide in the post-implantation embryo culture. *Toxicologist 12(1)*, 333.

Haberstroh, K. M. and Kapron, C. M. (1998). Cellular mechanisms of cadmium embryonic toxicity. *Teratology 57(4/5)*, 228.

Laschinski, G., Vogel, R. and Spielmann, H. (1989). Establishment of embryonic stem cells as a test system for embryotoxicity. *Teratology 40(3)*, 289.

Mayura, K., Smith, E. E., Clement, B. A. and Phillips, T. D. (1991). Evaluation of the developmental toxicity of chlorinated phenols utilizing Hydra attenuata and postimplantation rat embryos in culture. *Toxicol. Appl. Pharmacol. 108(2)*, 253-266.

McMillan, M. J., Ou, Y. C., Ponce, R. A. et al. (1997). Investigation of the temporal correlation between methylmercury induced cell cycle arrest and indution of cell cycle specific genes. *Teratology 55(1)*, 60.

Mirkles, P. (1996). Prospects for the development of validated screening tests that measure developmental toxicity potential: view of one skeptic. *Teratology 53(6)*, 334-338.

Mummery, C. L., Slager, H. G., van Inzen, W. et al. (1993). Regulation of growth and differentiation in early development: of mice and models. *Reprod. Toxicol. 7(1)*, 145-154.

NICEATM (National Toxicology Program Interagency Center for the Evaluation of Alternatives Toxicological Methods) (2000). Background Review Document. Frog Embryo Teratogenesis Assay – Xenopus (FETAX). March 10.

Newall, D. R. and Beedles, K. E. (1994). The stem-cell test: A novel in vitro assay for teratogenic potential. *Toxic.*

In Vitro 8(4), 697-701.
OECD (Organization for Economic Cooperation and Development). (1981). OECD guideline for testing of chemicals 414: Teratogenicity. OECD, Paris.
OECD (1983a). OECD guideline for testing of chemicals 415: One-generation reproduction toxicity study. OECD, Paris.
OECD (1983b). OECD guideline for testing of chemicals 416: Two-generation reproduction toxicity study. OECD, Paris.
OECD (1995). OECD guideline for testing of chemicals 421: Reproduction/developmental toxicity screening test. OECD, Paris.
OECD (1996). OECD guideline for testing of chemicals 422: Combined repeated dose toxicity study with the reproduction/developmental toxicity screening test. OECD, Paris.
Prieto, P. (2000). ECVAM's in-house prevalidation/validation studies in the areas of haematotoxicity, reproductive toxicity, metabolism-mediated toxicity and epithelial barrier function. *Sci. Total. Environ. 247*, 349-354.
Renault, J., Caillaud, J., Melcion, C. and Cordier, A. (1989). Micromass limb bud cell culture: a model for studying teratogenic mechanisms. *Teratology 40(3)*, 295.
Renault, J., Caillaud, J. and Chevalier, J. (1995). Ultrastructural characterization of normal and abnormal chondrogenesis in micromass rat embryo limb bud cell cultures. *Toxicol. Appl. Pharmacol. 130(2)*, 177-187.
Scholz, G., Pohl, I., Genschow, E. et al. (1999a). Embryotoxicity screening using embryonic stem cells in vitro: Correlation to in vivo teratogenicity. *Cells Tissues Organs 165(3-4)*, 203-211.
Scholz, G., Genschow, E., Pohl, I. et al. (1999b). Prevalidation of the embryonic stem cell test (EST). A new in vitro embryotoxicity test. *Toxic. In Vitro 13(4-5)*, 675-681.
Schwetz, B. A. (1993). Utility of in vitro assays: summary of international workshop on in vitro methods in reproductive toxicity. *Reprod. Toxicol. 7*, 171-173.
Spielmann, H., Pohl, I., Dsoring, B. et al. (1997). The embryonic stem cell test, an in vitro embryotoxicity test using two permanent mouse cell lines: 3T3 fibroblasts and embryonic stem cells. *In Vitro Toxicol. 10(1)*, 119-127.
Spielmann, H., Scholz, G., Bremer, S. and Holzhuetter, H. G. (1998). In vitro testing of embryotoxicity using the embryonic stem cell test (EST): an interlaboratory comparison. *In Vitro Cell Dev. Biol. - Animal. 34 (3 Pt 2)*, 11A.
USEPA (United States Environmental Protection Agency) (1998a). Health Effects Test Guidelines. OPPTS 870.3770. Prenatal Developmental Toxicity Study.
USEPA (1998b). Health Effects Test Guidelines. OPPTS 870.3800. Reproduction and Fertility Effects.
USEPA (2000a). Health Effects Test Guidelines. OPPTS 870.3550. Reproduction and Developmental Toxicity Screening Test.
USEPA (2000b). Health Effects Test Guidelines. OPPTS 870.3650. Combined Repeated Dose Toxicity Study with Reproduction and Developmental Toxicity Screening Test.
Walum, E. and Flint, O. P. (1993). Selective effects of acrylamide, methylene bisacrylamide, and haloperidol on neuronal development in rat embryo midbrain micromass cultures. *In Vitro Toxicol. 6(2)*, 125-134.
Watanabe, T. and Iwase, T. (1996). Developmental and dysmorphogeneic effects of glufosinate ammonium on mouse embryos in culture. *Teratog. Carcinog. Mutagen. 16(6)*, 287-299.
Whittaker, S., Wroble, J., Silbernagel, S. and Faustman, E. (1993). Characterization of cytoskeletal and neuronal markers in micromass cultures of rat embryonic midbrain cells. *Cell Biol. Toxicol. 9(4)*, 359-375.

6 Genetic engineering methodologies

6.1 Background

To say that genetic engineering is a controversial issue is to make an understatement. The controversy stems from the moral debate over whether the tremendous contribution "engineered" genes can make to medical science is justified in view of the possibility of inappropriate, misguided use of this relatively new technology. Genetic engineering can indeed be used to enhance the welfare of both humans and animals. However, questions remain as to whether humane considerations can be incorporated into the medical research needed to develop applications of genetic engineering. Moreover, there are concerns over whether the application itself is, in some instances, humane.

Examples of recent events in the realm of genetic engineering include the patenting of a process for engineering human cells that can be used to grow a variety of human tissues in the laboratory (Anonymous, 1998). Researchers claim that the techniques could be used to grow an endless supply of human organs for transplant. Furthermore, human genes have been successfully inserted into a pair of lambs supposedly endowing them with the DNA to aid burn victims (Arent, 1999). Moreover, the American Medical Association recently listened to proposed ethical guidelines for animal-to-human organ or cell transplants. The committee is recommending, among other things, that human subjects in clinical trials of animal-to-human transplantation – or xenotransplantation – be required to waive their traditional right to drop out of a study. This suggestion has met with considerable resistance from groups such as the Campaign for Responsible Transplantation (Philipkoski, 2000). Hence, genetic engineering continues to find numerous applications in the field of biomedical research and to provoke heated debate over the

ethical implications of those applications.

What specific contributions can genetic engineering make to the animal welfare movement? How are animals adversely affected by genetic engineering? Quite possibly, answers will be found as genetic engineering evolves. However, preliminary consideration of the issues will contribute much to our preparedness.

Such consideration has been the focus of several meetings convened under the auspices of the European Center for the Validation of Alternative Methods (ECVAM). One workshop on genetic engineering was organized in "recognition that genetically engineered cell lines, particularly those expressing xenobiotic metabolizing enzymes, are being used increasingly in toxicological and pharmacological studies" (Weibel et al., 1997). An additional ECVAM gathering on the use of transgenic animals in the European Union was held to "formulate a set of guidelines to assist regulatory authorities in the EU in deciding whether to permit and/or how to regulate research involving transgenic animals". (Metpham et al., 1998). Much work is still needed, however, to ensure that the benefits of genetic engineering can be realized without sacrificing animal welfare.

6.2 Genetically engineered cell lines

Genetically engineered cells can be used for determining the activity of specific enzymes in the metabolic profiles of xenobiotics, to screen chemicals for toxicity, and to evaluate molecular mechanisms of action. Heretofore, although representing a useful alternative test system for cytotoxicity or genotoxicity, established cells lines lacked sufficient endogenous xenobiotic metabolizing enzyme activity to make them a good predictor of human or animal response to toxicants. However, Weibel et al. (1997) indicate that xenobiotic metabolizing enzymes have now been expressed in bacteria, yeast, insect cells, and mammalian cells. They cite 20 examples of cell lines, mostly derived from human and hamster tissues, which have been genetically engineered for expression of xenobiotic metabolizing enzymes in the last decade. They also suggest that human cells expressing human xenobiotic metabolizing enzymes are more likely to be predictive of human susceptibility to toxicants than non-human cells.

Crespi has written extensively about the use of genetically engineered human cells. He suggests that, "Identification of the human enzymes involved in the metabolism of a series of investigational compounds (e.g. candidates for drug development) can help direct the choice, among the alternatives, of the best investigational compound to take into development, can help focus *in vivo*

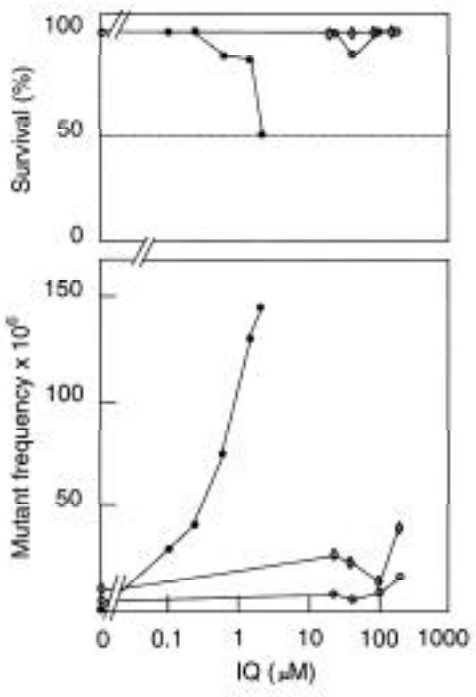

Fig. 11: Mutagenic activation of IQ in the cell lines expressing CYP1A2 and NAT.
ANM-13 and ANP-25 are cell lines expressing CYP1A2 plus NAT or NAT2, respectively. Top, relative survival rates after a 24-h treatment with IQ. Bottom, mutant frequency expressed as the number of 6-thio-guanine-resistant cells/10^6 viable cells (from Yanagawa et al., 1994).
o: A2R-5; ◊: ANM-13, ●: ANP-25

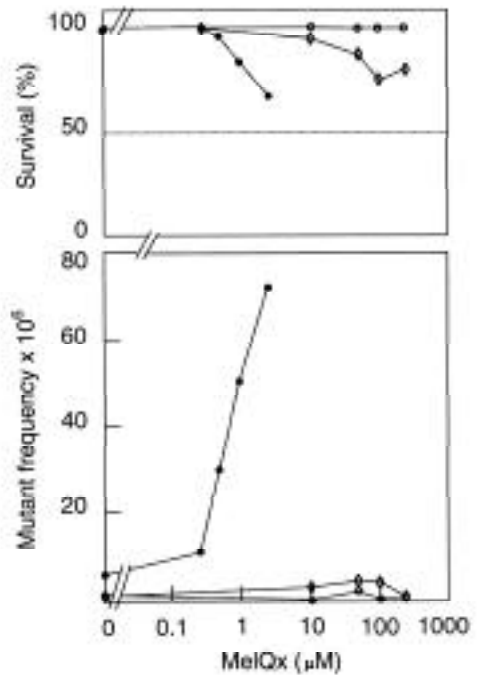

Fig. 12: Mutagenic activation of MeIQx in the cell lines expressing CYP1A2 and NAT (from Yanagawa et al., 1994).
o: A2R-5; ◊: ANM-13, ●: ANP-25

studies to specific areas of concern and can help elucidate the mechanisms for interindividual variability in response or toxicity" (Crespi, 1995). He also comments that much of the progress that has been made in the development of cell culture systems that contain xenobiotic-metabolising capacities is due to recombinant DNA techniques. Complementary DNA (cDNA) encoding nearly every major human P450 form has been isolated and sequenced.

Crespi and Penman (1997) have studied the use of single P450 enzyme systems, relative to multienzyme systems using cDNA. They suggest that this approach has major advantages and disadvantages. Direct incubation of a test compound with microsomes prepared from cells expressing a single cytochrome P450 form coupled with an analysis of metabolite formation or loss of parent compound can yield significant insight into cytochrome P450 "form-selective" metabolism of a test compound. Use of single P450 enzyme systems can also be used to study the inhibition of human, drug-metabolizing cytochromes P450. In a recent study, microtiter plate-based fluorometric assays for the activities of five principal drug-metabolizing enzymes, CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4 were evaluated. (Crespi et al., 1997). The authors found that their observed IC_{50} values were in good agreement with those reported for the same enzyme/inhibitor pairs reported by others.

Single-enzyme experiments may not predict, however, the activity of the balance of enzymes present in human liver microsomes. In their review of genetic engineering, Weibel et al. (1997) recommended that, "For studies on mechanisms of metabolic activation, enzymes acting in concert, such as cytochrome P4501A2 and N-acetyltransferases or sulphotransferases, need to be co-expressed in recombinant systems. This also applies to enzymes involved in competing metabolic pathways". Crespi (1995) concurs by suggesting that, "co-expression of many enzymes within a single cell will be an important technique for the analysis of the balance of Phase I and Phase II enzymes on overall metabolism and/or toxicity of xenobiotics". He and others demonstrated the feasibility of this approach with the development of MCL-5 cells that coexpress five cDNAs (Crespi et al., 1991). More recently, Sawada and Kamataki (1998) have reviewed the stable expression of cytochrome P450 in mammalian cell lines and its application to toxicity testing. In their discussion the successful reconstruction of an entire metabolic activation system for certain heterocyclic amines is described. The system consists of mammalian cell lines that simultaneously express a form of human P450 and a phase II enzyme, N-acetyltransferase (Fig. 11 and 12) (Yanagawa et al., 1994).

Kranendonk et al. (1999) have also recently reported on the genetic engineering of four new *Escherichia coli* tester bacteria, coexpressing human CYP1A1, CYP2A6, CYP3A4 or CYP3A5 with human NADPH cytochrome P450 reductase (RED) by a biplasmid co-expression system, recently developed to express human CYP1A2 in the tester strain MTC. They found these four new strains to "contain stable CYP- and RED-expression, significant CYP-activities and demonstrated significant bioactivation activities with several diagnostic carcinogens" (Tab. 13).

Tab. 13: Mutagenic activities of chemical carcinogens in strain MTC/CYP expressing five different human CYP isoforms (from Kranendonk et al., 1999)

Carcinogen	Strain	Mutagenic activity (revertants per nanomole ± standard deviation)
2AA	MTC1A2	3433 ± 447
	MTC^2+S9	108 ± 4[a]
AFB1	MTC1A1	77 ± 12
	MTC1A2	5437 ± 555
	MTC2A6	421 ± 10
	MTC3A4	2349 ± 330
	MTC3A5	978 ± 59
	MTC^2+S9	13247 ± 732[a]
B[a]P	MTC1A1	798 ± 60
	MTC^2+S9	84 ± 8[a]
DMBA	MTC1A1	682 ± 77
	MTC^2+S9	5 ± 1[b]

[a] Determined with non-induced strain MTC1A2

[b] Determined with strain MX100

However, Crespi and Penman (1998) have cautioned that considerable work remains to be done to develop appropriate approaches for use of these data to predict the consequences of human exposure to xenobiotics. Both appropriate validation of systems containing human, cDNA-expressed xenobiotic metabolizing enzymes and integration of experimental results from host cells expressing different amounts of enzyme activity than those found *in vivo* need to be addressed.

Other areas of research benefiting from human cells genetically engineered to contain cytochrome P450 metabolic enzymes include cancer research and therapy. Recent examples include the work of Hanna et al. (2000) who researched the association of cytochrome P450 1B1 (CYP1B1) polymorphisms with functional differences in estrogen hydroxylation activity. They evaluated whether inherited variants of CYP1B1 differ from wild-type CYP1B1 in estrogen hydroxylase activity by developing recombinant wild-type and five polymorphic variants of CYP1B1. The CYP1B1 variants formed 4-OH-E2 as the main product (as did the wild type), but displayed 2.4- to 3.4-fold higher catalytic efficiencies. The variant enzymes also exceeded wild-type CYP1B1 with respect to 2- and 16alpha-hydroxylation activity. The authors concluded that inherited alterations in CYP1B1 estrogen hydroxylation activity may be associated

with significant changes in estrogen metabolism and, "thereby, may possibly explain interindividual differences in breast cancer risk associated with estrogen-mediated carcinogenicity". Moreover, Waxman et al. (1999) have reviewed recent advances in cancer gene therapy in which an intratumoral prodrug activation strategy based on the combination of a cytochrome P450 gene with the gene encoding NADPH-P450 reductase is utilized. Transduction of tumors with a prodrug-activating P450 gene, followed by prodrug treatment, greatly increases intratumoral formation of activated drug metabolites. This leads to "more efficient killing of the transduced tumor cells without a significant increase in host toxicity".

Among the conclusions and recommendations of the ECVAM report (Weibel et al, 1997) regarding further development of recombinant systems were the following:

- There is a need to develop systems which express xenobiotic metabolizing enzymes, other than the cytochromes P450.

Recent work by Hu et al. (1999) studied the transport and metabolic characterization of Caco-2 cells expressing CYP3A4 and CYP3A4 plus oxidoreductase. They found that the monolayers of newly transfected cells (CYP3A4 + oxidoreductase) have significantly increased levels of CYP3A4 activities compared to untransfected cells. These cell monolayers also have desirable morphological and transport characteristics that are similar to untransfected cells (Tab. 14).

- For risk assessment purposes, xenobiotic metabolizing enzymes of commonly used species of laboratory animals need to be expressed in recombinant cell systems. Priority should be given to rat enzymes.

Recent work by Grove et al. (2000) investigated the influence of rat UGT1A7 on B[a]P-induced cytotoxicity. Human lymphoblastoid L3 cells were transfected with pMF6 (control expression vector), p167Dtk2 (microsomal epoxide hydrolase expression vector), or P167Dtk2-1A7 (epoxide hydrolase/UGT1A7 coexpression vector), and the cell populations were compared for sensitivity to B[a]P-induced effects. The data obtained suggested that UGT1A7 may be preferentially active toward B[a]P-quinones and that UGT1A7 may represent the PAH-inducible UGT activity previously implicated in protection against toxic redox cycling by B[a]P-3,6-quinone.

- It is essential to establish mammalian cell lines which exhibit high levels of enzyme activities (for example, of the magnitude achievable in bacterial recombinant systems). Preferably, the host cells should be of human origin.

Crespi (1995) echoed these sentiments suggesting that "higher expression levels are necessary for routine use of such a system for metabolite analysis".

Recent work by Scott et al. (1999) studied the expression and partial purification of a recombinant secretory form of human liver carboxylesterase (CaE) from human cells in which the carboxylesterase was overexpressed. Carboxylesterases play an important role in the detoxification of xenobiotic chemicals that contain organophosphate compounds; high activity is found in the liver. In Scott's work, a single amino acid change was made in the C-terminal retrieval signal of a human liver CaE. A recombinant plasmid, pRc/CMV-mh-CaE, containing the amino acid change, was isolated and stably integrated into human 293T cells. Expression of the altered cDNA resulted in secretion of an active CaE up to levels of 500 enzyme units per liter of growth medium. This secretory CaE displayed isoelectric focusing patterns similar to those of the native enzyme with no observable changes in activity. The secreted enzyme was partially purified and CaE retained a high level of enzymatic activity.

All of these research findings indicate that progress is steadily being made in recombinant genetic engineering as it relates to toxicology and drug metabolism as well as to areas such as cancer research.

6.3 Transgenic animals

Transgenic animals are a product of genetic engineering and they are a good example of the dichotomy inherent to this area of research as mentioned earlier in this article. There are definitely uses for transgenic animals that contribute a great deal to toxicological research. However, their negative effects on animal welfare may outweigh their contributions. Metpham et al. (1998) state that "while transgenic animals might allow reduction and refinement in animal

Tab. 14: Metabolic properties of model compounds in Caco-2 cell variants (from Hu et al., 1999)

	Metabolic rates (pmol/min/mg protein)		
Parameters/drugs	**TC7 clone**	**CYP3A4#4**	**CYP3A4+OR#8**
Drugs			
testosterone (Tes)	BDL[a]	7-32	5.20 ± 0.48
nifedipine (Nif)	1	2-10	1.36 ± 0.08
linopirdine (Lin)	BDL	12.3 ± 1.4	ND
midazolam	ND	1.09 ± 0.12	ND
warfarin	BDL	BDL	ND
Drug + Modulator (% Control)			
Tes + ketoconazole	BDL	7.0 ± 0.0	ND
Nif + ketoconazole	12 ± .16	5 ± 0.2	ND
Nif + TPA	ND[b]	145 ± 5	ND
Lin + ketoconazole	ND	57.6 ± 3.8	ND
Lin + TPA	ND	ND	4.71 ± 0.07[c]
Tes + TPA	ND	183 ± 21	7.57 ± 0.44[c]

[a] "BDL" stands for "below detection limit" of 0.25 pmol/min/mg protein.

[b] "ND" stands for "not determined".

[c] Actual rates of metabolism with a unit of pmol/min/mg protein.

use via more precise gene targeting in breeding programs, these objectives are threatened by transgenic procedures which could promote greater animal use, a greater variety of applications, and an increased likelihood of animal suffering".

Currently transgenic animals are used in numerous basic research applications, as well as in the development of desirable traits in agricultural products and livestock. They are considered "bioreactors" for the mass production of biological products. They are also used to test vaccines. Moreover, several types of transgenic animals are either used or "envisaged" for specific toxicological testing procedures.

One of the most widely discussed uses of transgenesis in toxicology is that of transgenic mice as predictive models for identifying carcinogens. Advantages to using this model or others like it include use of fewer animals, performance of the assay in a short time period, and much lower cost than the average two-year rodent bioassay. Yamamoto et al. (1998) states that, "*In vitro* testings, i.e. mutagenicity tests, transformation assays, etc. may be helpful to predict carcinogenicity; but such *in vitro* testing is still subsidiary. Carcinogenicity testing using experimental animals is the only way for the prospective identification of possible human carcinogens". They indicate that animals susceptible to carcinogens (i.e transgenic (Tg) animals) are indispensable and describe the validation of transgenic mice carrying the human prototype c-Ha-*ras* gene as a bioassay model for rapid carcinogenicity (Tab. 15).

Others have also evaluated transgenic mice as predictive models for carcinogenesis. Eastin (1998) describes a preliminary assessment by the U.S. National Toxicology Program (NTP) of the utility of the $p53^{def}$ (hemizygous for the tumor-suppressor gene) and the Tg.AC (carrier of an activated H-*ras* oncogene) transgenic mice strains as models for identifying chemical carcinogens and assessing risk. Moreover, Contrera and DeGeorge (1998) as well as Monro and MacDonald (1998) discuss the evaluation of the carcinogenic potential of pharmaceuticals using transgenic mice in response to recommendations by the International Conference on Harmonisation (ICH) that allow the substitution of one species in the standard 2-year cancer bioassay with a short- or medium-term alternative model (ICH, 1996).

Furthermore, Thompson et al. (2000) have recently investigated a battery of prescreening methods based on comparisons with the Tg.AC mouse model. They developed a battery based on correct responses to 24 compounds tested previously in Tg.AC mice. Included in the battery were assays chosen to reflect molecular pathways possibly involved in the Tg.AC papilloma formation: zeta-globin promoter-luciferase construct stably expressed in K562 cells (Zeta-Luc) and three of the stress-response element-chloramphenicol acetyltransferase (CAT) fusion constructs stably expressed in HepG2 cells that are a part of the CAT-Tox (L)iver assay. Stress response elements chosen were the c-fos promoter, the gadd153 promoter, and p53 response element repeats. The gadd153-CAT assay showed the strongest correlation with activity in the Tg-AC assay. This correlation was improved further by adding the Zeta-Luc assay as a second-stage screen. Hence, efforts to augment the Tg.AC mouse model with other assays appear promising.

Use of the transgenic mouse bioassay for carcinogenicity has been criticized by advocates of *in vitro* bioassays, however. For example, Van Zeller and Combes (1999) evaluated an interlaboratory collaborative study organized by the International Life Sciences Institute (ILSI) that included four transgenic carcinogenicity assays in mice (the c-Ha-*ras*, Tg.AC, P53+/- and XPA systems). They concluded that, "to date, the data suggest that none of these assays is appropriate for inclusion in a carcinogenicity testing strategy". They suggested that, "more emphasis should be placed on developing replacement alternative assays which are capable of identifying and characterizing carcinogens of human relevance". The Syrian hamster embryo cell transformation assay was considered to represent perhaps a more appropriate assay for evaluation of human carcinogenesis by these researchers.

Additional criticism has been directed at transgenic models for prion disease (Jenkins and Combes, 1999). Prion diseases are fatal neurological disorders caused by proteinaceous agents known as prions. An example of this type of disease is that of Creutzfeldt Macomb disease in humans. These researchers feel that the use of transgenic mice to study prions should be augmented by *in vitro* alternatives. Examples of these alternatives include cell-free conversion assays, cell culture systems, immunoassays for the pathogenic form of the prion protein, and perhaps yeast-based systems because prion-like proteins have been identified in yeast.

These criticisms of transgenic animal models reflect a growing concern about greater use of transgenic animal models without appropriate consideration of *in vitro* alternatives which may, in fact, be as predictive, as the transgenic models. In response to these and other animal welfare issues about transgenesis, Delpire et al. (1999) have proposed a new ethical scheme addressing the use of laboratory animals for biomedical purposes. They suggest that current European Union legislation relating to the use of laboratory animals was implemented before transgenesis and its applications were appropriately researched. They recommend that "a specific cluster of questions dealing with transgenic animals" could be incorporated at the European level when reviews of project proposals involving animal experimentation occur.

What specifically can go wrong with transgenic animals? Van Reenen and Blockhuis (1993, 1997) present three general factors that can result in negative effects in transgenic animals: reproductive and other biotechnological interventions, mutations, and expression of the transgene. Metpham et al. (1998) discuss these factors in-depth. They suggest that *in vitro* procedures employed both before and after microinjection can lead to increased gestation length, body weight, incidence of dystocia, and perinatal loss and anomalies as compared to *in vivo* procedures. Furthermore, foreign DNA often becomes integrated within or adjacent to an endogenous gene creating a new mutation that has deleterious effects on the host. Finally, transgenic animals may express characteristics that are harmful to them. The degree of harm ex-

Tab. 15: Results of rapid carcinogenicity testing using rasH2 mice (from Yamamoto et al., 1998)

Chemicals tested	Dose	Route of administration	Rapid tumor response[a] in Tg mice	Tumor incidence[b]	Malignant tumors	
					Tg	Non-Tg
Mutagenic *(Salmonella)* carionogens						
p-Cresidine	0.25 or 0.5% for 26 wk	Feed	+*	Tg Non-Tg	+	+
Cyclophosphamide[c]	30mg/kg x 2/wk for 25 wk	Gavage	+/-	Tg Non-Tg	+	-
DEN[c]	90mg/kg x 1	Intraperitoneal	+*	Tg > Non-Tg*	+	-
1,2-Dimethylhydrazine[g]	20mg/kg x 1/wk for 20 wk	Subcutaneous	+*	Tg > Non-Tg*	+	-
4HAQO[e]	10 or 20mg/kg x 1	Intravenous	+*	Tg > Non-Tg*	++	+
MAM[c]	20mg/kg x 1/wk for 6 wk	Subcutaneous	+	Tg > Non-Tg	+	-
Melphalan	0.3 or 1.5mg/kg x 1/wk for 26 wk	Intraperitoneal	+	Tg > Non-Tg	+	-
MNNG[c]	2.5mg x 1	Gavage	+*	Tg > Non-Tg*	+	-
MNU[c]	75mg/kg x 1 or 15mg/kg x 5	Intraperitoneal	+*	Tg > Non-Tg*	+	-
NNK[e]	3 or 6mg x 1/wk for 2 wk	Intraperitoneal	+*	Tg < Non-Tg*	+	-
4NQO[c]	15mg/kg x 1	Subcutaneous	+*	Tg > Non-Tg*	+	-
Phenacetin[e]	0.7 or 1.4% for 24 wk	Feed	+*	Tg > Non-Tg*	+	-
Procarbazine[e]	6 or 12mg/kg x 3/wk for 24 wk	Intraperitoneal	+	Tg > Non-Tg	+	-
4,4'-Thiodianiline[e]	2000 or 4000 ppm for 24 wk	Feed	+*	Tg > Non-Tg*	+	-
Thiotepa[e]	1 or 2mg/kg x 3/wk for 24 wk	Intraperitoneal	+*	Tg > Non-Tg*	+	-
Vinyl carbamate[f]	60mg/kg x 1	Intraperitoneal	+*	Tg > Non-Tg*	++	+
4-Vinyl-1-cyclohexene diepoxide[e]	5 or 10mg x 5/wk for 24 wk	Dermal	+*	Tg > Non-Tg*	++	+
Nonmutagenic *(Salmonella)* carcinogens						
Benzene[e]	50 or 100 mg/kg x 5/wk for 24 wk	Gavage	+*	Tg > Non-Tg*	+	-
Cyclosporin[e]	10 or 25 mg/kg x 5/wk for 24 wk	Gavage	+	Tg > Non-Tg	+	-
1,4-Dioxane[e]	0.5 or 1% for 24 wk	Water	+	Tg > Non-Tg*	+	-
Ethyl acrylate[e]	100 or 200 mg/kg x 5/wk for 24 wk	Gavage	+*	Tg > Non-Tg*	+	-
Ethylene thiourea[d]	0.3% for 28 wk	Feed	+*	Tg Non-Tg	+	+
1,1,2-Trichloroethane[e]	100 or 200mg/kg	Gavage	-			
Mutagenic *(Salmonella)* noncarcinogens						
p-Anisidine	0.225 or 0.45% for 26 wk	Feed	-			
8-Hydroxyquinoline[e]	1500 or 3000 ppm for 24 wk	Feed	-			
4-Nitro-*o*-phenylenediamine[e]	3750 or 7500 ppm for 24 wk	Feed	+/-	Tg Non-Tg	-	-
Nonmutagenic *(Salmonella)* noncarcinogenS						
Resorcinol	225mg/kg x 5/wk for 26 wk	Gavage	-			
Rotenone[e]	600 or 1200 ppm for 24 wk	Feed	-			
Xylene (mixed)	500 or 1000 mg/kg x 5 /wk for 24 wk	Gavage				

[a] Rapid tumor response means that tumors actually developed within 26 weeks, (28 weeks for ethylen thiourea) in *ras*H2 (Tg) mice in response to tested chemicals.

(+*) "Positive" indicates that the incidence of at least one type of tumor developed in chemical-treated Tg mice was significantly ($p<0.05$; Fisher´s exact test) higher than that in control vehicle-treated Tg mice.

(+) indicates that the incidence of at least one type of tumor developed in chemical-treated Tg mice was ≥25% but not statistically significant versus control.

(+/-) "Positive-negative" indicates that the incidence of at least one type of tumor developed in chemical treated Tg mice was more than 13% but less than 25%.

(-) "Negative" indicates that the incidence of at least one type of tumor developed in chemical treated Tg mice was ≤13% or that tumor incidence was similar between the vehicle-treated control and dosed groups.

[b] Tumor incidence: Incidence of at least one type of tumor (in cases of p-cresidine and ethylene thiourea, incidences of diffuse hyperplasia of urinary bladder and thyroid follicular cell hyperplasia, respectively) developed in chemical-treated Tg mice was statistically different from that of corresponding non-Tg mice ($p<0.05$; Fisher´s exact test)

References of the original publication
[c] Data from Yamamoto et al (5).
[d] Data from Yamamoto et al. (6).
[e] Data from Urano et al. (7).
[f] Data from Mitsumori et al (8).
[g] Data from Itoh et al. (unpublished data).

pressed was noted to be dependent on the biological properties of the resulting protein, the tissue in which transgenes are expressed, the route of secretion of the gene product, and the level of transgene expression.

In direct contrast to concerns that the harmful effects of transgenic expression be avoided, are transgenic models of disease. These animals are used as tools in which the harmful effects of transgenic expression are actively sought. As such they represent an area of transgenic research in which the moral dilemma between animal welfare and biomedical advance can perhaps be most clearly observed.

That research involving transgenic models of disease is proliferating is certainly true (for example, Ware et al., 2000 (Bernard-Soulier syndrome); Shirai et al., 2000 (prostate carcinogenesis); Bardeesy et al., 2000 (melanoma); Guyer, 2000 (neurodegenerative diseases)). Unfortunately, the current focus of much of this research seems to be generation of a transgenic model without much consideration for the welfare of the animals involved. Metpham suggests, however, that "…it should be possible to generate animals that have a number of salient features of the disease but do not develop the full disease condition". This suggestion should be achievable in view of the enhanced capability with which geneticist's can accurately identify causative genes in the disease process. Hardouin and Nagy (2000) in their review of mouse models for human disease cite the fact that "In the last two decades by the advent of novel and very efficient genetic tools, the creation of disease models is significantly advancing. We have reached the stage where mouse geneticists feel there is no limitation in creating phenocopies (or genocopies) of any mutations or chromosomal aberrations identified in disease". Hopefully, this capability will result in production of more humane transgenic disease models.

More significant perhaps than the effects of transgene expression of any kind, are the more general, but perhaps less obvious, effects of introduction of transgenes into a host organism. Van der Meer and colleagues (1999), in an important attempt to differentiate between the effects of the technique of transgenesis and the effects caused by the expression of the transgene, studied behavioral responses and morphological development of newborn mice. Newborn animals were subjected to various behavioral tests and their growth and morphological characteristics were measured from birth up to the age of 3 weeks. The results indicated that the microinjected DNA-construct influenced the survival rate during the first 2-3 days after birth. The average loss of pups was about 10%, in contrast to the groups without the DNA construct, in which none of the pups died. Body weight gain was significantly lower for the DNA-construct pups, but only during the first 11 days. No significant differences in morphological characteristics or behavioral development were noted. The authors suggested that, "Before general conclusions about the extent to which the technique of transgenesis affects animal welfare can be drawn, more transgenic lines should be studied in this way." Evaluations such as this one should indeed be performed if we are to draw viable conclusions from data generated from any transgenic animal.

6.4 Summary, conclusions, and future work

Genetic engineering offers tremendous opportunities for the advancement of biomedical research in many areas, including toxicology. The expression of metabolic enzymes in human cell culture systems represents a major toxicological tool with which to study drug and/or toxicant metabolism and elicit a clearer understanding of many mechanisms of action. Current research trends are toward the isolation of the activity of single enzymes in a metabolic profile followed by study of the activities of multiple enzymes acting in sequence as is the case *in vivo*. Considerable progress has been made thus far. However, caveats have been issued with regard to the extent to which these systems can be reasonably compared to *in vivo* situations. Appropriate validation of much of this work remains to be accomplished.

The generation and use of transgenic animals to study questions of biomedical interest have been questioned by many in view of the moral and ethical dilemmas presented by these activities. That transgenic animals may contribute to the reduction of animal use in toxicological experiments, particularly studies of carcinogenicity, is not disputed. However, advocates of replacement alternatives argue that *in vitro* alternatives to this type of toxicity testing have not been given adequate attention. Moreover, although transgenic models of disease offer a potent method with which to study disease conditions, animal welfare is often not a foremost consideration in many of these research endeavors.

Despite the controversy surrounding use of transgenic animals, recent recommendations by the ICH have indicated that a transgenic model can be considered as a replacement for one species in chronic carcinogenic evaluations of pharmaceutical products. Several such models are undergoing evaluation and battery-type approaches incorporating more than one model have been suggested. Many individuals however, remain circumspect of the use of transgenic animals in general for humane reasons. Actions such as those in which a "specific cluster of questions dealing with transgenic animals" is incorporated into any review of research proposals in which transgenic models are used should be encouraged.

Moreover, the blatant disregard for animal welfare that is often inherent in any experimental rush to define the biochemical lesion resulting in a particular disease should be discouraged. As suggested by Metpham above, animal models of disease should be designed so that the salient features of the disease are emphasized and any unnecessary untoward effects on the animal minimized. Given the very sophisticated genetic tools we have on hand these days, this should not be an insurmountable, or even difficult, task.

References

Anonymous (1998). Human cell growth patented. http://www.wired.com/news/technology/0,1282,16730l,00.html

Arent, L. (1999). Lambs get human genes. http://www.wired.com/technology/0,1282,20874,00.html

Bardeesy, N., Wong, K., DePinho, R. and Chin, L. (2000). Animal models of melanoma: recent advances and future prospects. *Adv. Cancer Res. 79*, 123-156.

Contrera, J. and DeGeorge, J. (1998). In vivo transgenic bioassays and assessment of the carcinogenic potential of pharmaceuticals. *Environ. Hlth. Perspect. 106(1)*, 71-80.

Crespi, C., Gonzalez, F., Steimel, D. et al. (1991). A metabolically competent human cell line expressing five complementary DNAs encoding procarcinogen-activating enzymes: Application to mutagenicity testing. *Chem. Res. Toxicol. 4*, 566-572.

Crespi, C. (1995). Xenobiotic-metabolizing human cells as tools for pharmacological and toxicological research. *Adv. Drug Res. 26*, 180-235.

Crespi, C. and Penman, B. (1997). Use of cDNA-expressed human cytochrome P450 enzymes to study potential drug-drug interactions. *Adv. Pharmacol. 43*, 171-188.

Crespi, C., Miller, V. and Penman, B. (1997). Microtiter plate assays for inhibition of human, drug-metabolising cytochromes P450. *Anal. Biochem. 248*, 188-190.

Crespi, C. and Penman, B. (1998). Use of human engineered cells in toxicology. *Hum. Exp. Toxicol. 17(9)*, 485.

Delpire, V., Mepham, T. and Balls, M. (1999). A proposal for a new ethical scheme addressing the use of laboratory animals for biomedical purposes. *ATLA 27*, 869-881.

Eastin, W. (1998). The U.S. National Toxicology Program evaluation of transgenic mice as predictive models for identifying carcinogens. *Environ. Hlth. Perspect. 106 (Suppl. 1)*, 81-84.

Grove, A., Llewellyn, G., Kessler, F. et al. (2000). Differential protection by rat UDP-glucuronosyltransferase 1A7 against Benzo[a]pyrene-3,6- quinone versus Benzo[a]pyrene-induced cytotoxic effects in human lymphoblastoic cells. *Toxicol. Appl. Pharmacol. 162(1)*, 34-43.

Guyer, C. (2000). Recent developments in animal models of human neurodegenerative diseases. *Toxicol. Pathol. 28(2)*, 363-366.

Hanna, I. H., Dawling, S., Roodi, N. et al. (2000). Cytochrome P450 1B1 (CYP1B1) pharmacogenetics: association of polymorphisms with functional differences in estrogen hydroxylation activity. *Cancer Res. 60(13)*, 3440-3444.

Hardouin, S. and Nagy, N. (2000). Mouse models for human disease. *Clin. Genet. 57*, 237-244.

Hu, M., Li, Y., Davitt, C. et al. (1999). Transport and metabolic characterization of Caco-2 cells expressing CYP3A4 and CYP3A4 plus oxidoreductase. *Pharm. Res. 16(9)*, 1352-1359.

ICH (International Conference on Harmonisation) Expert Safety Working Group (1996). Testing for carcinogenicity of pharmaceuticals. *Federal Register 61*, 43298-43300.

Jenkins, E. and Combes, R. (1999). Transgenic models for prion disease: Have they outlived their useful purpose? *ATLA 27*, 827-838.

Kranendonk, M., Carreira, F., Theisen, P. et al.(1999). Escherichia coli MTC, a human NADPH P450 reductase competent mutagenicity tester strain for the expression of human cytochrome P450 isoforms 1A1, 1A2, 2A6, 3A4, or 3A5: catalytic activities and mutagenicity studies. *Mutat. Res. 441(1)*, 73-83.

Metpham, T., Coombes, R., Balls, M. et al. (1998). The use of transgenic animals in the European Union. The Report and Recommendations of ECVAM Workshop 28. *ATLA 26*, 21-43.

Monro, A. and MacDonald, J. (1998). Evaluation of the carcinogenic potential of pharmaceuticals. *Drug Safety 18(5)*, 309-319.

Philipkoski, K. (2000). The future: part man, part beast. http://www.wired.com/news/technology/0,1282,40512,00.html

Sawada, M. and Kamataki, T. (1998). Genetically engineered cells stably expressing cytochrome P450 and their application to mutagen assays. *Mutat. Res. 411(1)*, 19-43.

Scott, D., Chacko, T., Maxwell, D. et al. (1999). Expression and partial purification of a recombinant secretory form of human liver carboxylesterase. *Protein Expr. Purif. 17(1)*, 16-25.

Shirai, T., Takahashi, S., Cui, L. et al. (2000). Experimental prostate carcinogenesis – rodent models. *Mutat. Res. 462(2-3)*, 219-226.

Thompson, K., Rosenzweig, B., Tsong, Y. and Sistare, F. (2000). Evaluation of in vitro reporter gene induction assays for use in a rapid prescreen for compound selection to test specificity in the Tg-AC mouse short-term carcinogenicity assay. *Toxicol. Sci. 57(1)*, 43-53.

Van der Meer, M., Costa, P., Baumans, V., Olivier, B. and Van Zutphen, B. (1999). Welfare assessment of transgenic animals: behavioural responses and morphological development of newborn mice. *ATLA 27*, 857-868.

Van Reenen, C. and Blockhuis, H. (1993). Investigating welfare of dairy calves involved in genetic mode: problems and perspectives. *Livestock Production Science 36*, 81-90.

Van Reenen, C. and Blockbuis, H. (1997). Evaluation of welfare of transgenic farm animals: lessons from a case study in cattle. *Journal of the Royal Swedish Academy of Agriculture and Forestry. 136*, 99-105.

Van Zeller, A. and Combes, R. (1999). Transgenic mouse bioassays for carcinogenicity testing: A step in the right direction? *ATLA 27*, 839-846.

Ware, J., Russell, S. and Ruggeri, Z. (2000). Generation and rescue of a murine model of platelet dysfunction: the Bernard-Soulier syndrome. *Proc. Natl. Acad. Sci. USA 97(6)*, 2803-2808.

Waxman, D., Chen, L., Hecht, J. and Jounaidi, Y. (1999). Cytochrome P450-based cancer gene therapy: recent advances and future prospects. *Drug Metab. Rev. 31(2)*, 503-522.

Weibel, F., Andersson, T., Casciano, D. et al. (1997). Genetically engineered cell lines: Characterisation and applications in toxicity testing. The Report and Recommendations of ECVAM Workshop 26. *ATLA 25*, 625-639.

Yamamoto, S., Urano, K., Koizumi, H. et al. (1998). Validation of transgenic mice carrying the human prototype c-Ha-ras gene as a bioassay model for rapid carcinogenicity testing. *Environ. Hlth. Perspect. 106 (Suppl. 1)*, 57-69.

Yanagawa, Y., Sawada, M., Deguchi, T., Gonzalez, F. and Kamataki, T. (1994). Stable expression of human CYP1A2 and N-acetyltransferases in Chinese hamster CHL cells: mutagenic activation of 2-amino-3-methylimidazo[4,5-f]quinoline and 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline. *Cancer Res. 54*, 3422-3427.

7 Gene chip technology as an alternative to animal testing

7.1 Background

DNA microarrays, commonly referred to as "gene chips", form the basis of a new field of endeavor within toxicology, known as "toxicogenomics". The major issue to be addressed in toxicogenomics is the relevance of the endpoint(s) measured to the toxicological insult(s) studied. Questions concerning the relationship between the genetic changes observed on a microarray after exposure to a toxic chemical and the toxic insult produced at the biochemical, cellular, or tissue level remain to be answered. This type of question however, haunts most, if not all, techniques which are purported to be alternatives to the use of animals in toxicity testing. The answer is often found through analysis of biochemical mechanisms of toxic insult and often supported through exhaustive validation efforts that analyze the degree of correlation existing between *in vitro* alternatives and animal or human assays. In the case of gene chips, this analysis will have to be extended to analysis of genetic mechanisms of toxic insult. However, if reliable links between genetic changes and ensuing toxic responses can be made, and these genetic changes can be faithfully identified with gene chip technology, then gene chips would indeed represent an alternative method which would greatly reduce the number of, if not replace entirely, the animals used in these experiments.

7.2 What is a gene chip?

Gene chips have been defined with regard to both construction techniques and function. Ermolaeva et al. (1998) regarded gene chips as one of "several new technologies for studying the simultaneous expression of large numbers of genes". They divided these technologies into serial and parallel methods, and defined parallel approaches (such as microarrays) as those based on hybridization of mRNA or DNA to cDNA immobilized on glass or to synthetic oligonucleotides immobilized on silica wafers or "chips". Bowtell (1999) provides an expanded definition. His microarray is a "complete system" composed of three parts involving sample preparation ('the front end'), array generation and sample analysis ('middleware') and data handling and interpretation (the 'back end'). The 'front end' requires preparation of RNA from various sources, the 'middleware' is developed through target preparation, labeling and construction of the array, and the 'back end'is characterized by techniques for moving and handling the large quantities of data generated by these array experiments.

Brown and Botstein (1999) provide a detailed description of the development of cDNA microarrays. Arrays of discrete DNA sequences are printed on glass microscope slides using a robotic "arrayer" and the relative abundance of each gene sequence in two DNA or RNA samples isolated from two different cell populations is compared. The two samples are labeled with different fluorescent dyes and then mixed and hybridized with the arrayed DNA spots. Subsequently, fluorescence measurements are made with a microscope that illuminates each DNA spot and measures fluorescence for each dye separately. These measurements are then used to determine the ratio and, hence, the relative abundance, of the sequence of each specific gene in the two mRNA or DNA samples. These researchers indicate that although cDNA and oligonucleotide microarrays differ in detail, they share the same "essential simplicity of experimental design" which they attribute to the specificity and affinity of complementary base-pairing. Figure 13 contains a simplified diagram of sample preparation and hybridization to cDNA microarrays.

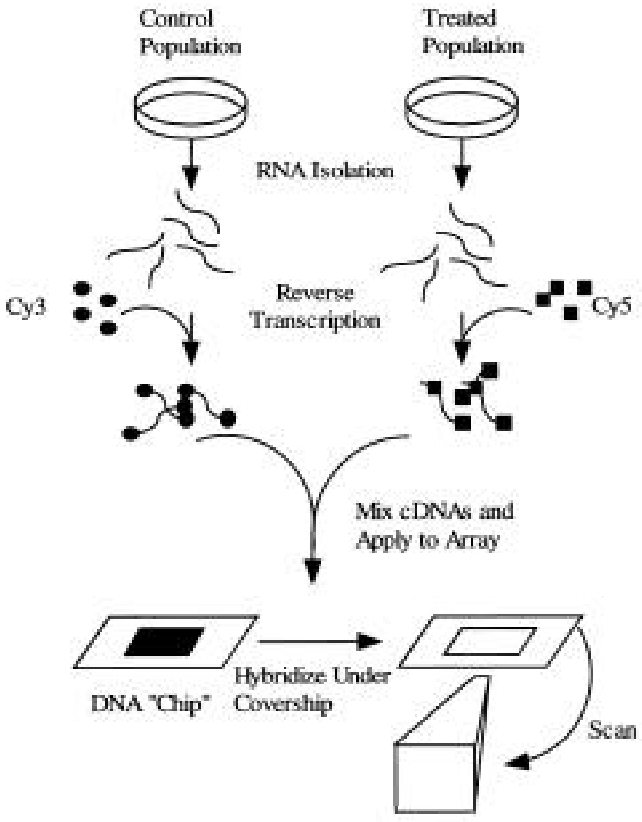

Fig. 13: Simplified overview of the method for sample preparation and hybridization to cDNA microarrays.
For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis (from Nuwaysir et al., 1999).

Lockhart et al. (1996), while acknowledging the advantages of cDNA spotting onto chips, pointed out that in order to monitor many genes, a large number of cDNAs or PCR products must be prepared, purified, quantitated, catalogued, and spotted onto a solid support. They describe development of an approach based on hybridization to small, high-density arrays containing thousands of synthetic oligonucleotides. These arrays are designed based on sequence information and are synthesized *in situ* using a combination of photolithography and oligonucleotide chemistry. The advantages to this approach are that oligonucleotide array construction is "direct, combinatorial synthesis of appropriate oligonucleotides based on sequence information alone eliminat[ing] the necessity of preparing and handling clones, PCR products, and cDNAs." Furthermore, they suggest that because oligonucleotide probes for each gene are specifically chosen and synthesized in known locations on the arrays, hybridization patterns and intensities can be interpreted with no additional sequencing or characterization.

Whether cDNA or oligonucleotide arrays, DNA-chip products are definitely benefiting from improvements in fabrication techniques that were developed initially for computer chip manufacture. As a result, organic structures (such as segments of DNA or mRNA) can now be applied to a substrate of inorganic materials. Inorganic substrates for DNA-chips are usually glass or plastic wafers. However, with the development of microassays that can be done on a chip, miniaturized glass tubes and reservoirs are also being utilized. Bonding of genetic sequences on a microchip using photolithographic processes as well as deposition of sequences onto the chip using a droplet sprayer or a robot are currently the principal methods used to join organic structures with inorganic substrates in chip manufacture (Hencke, 1998a).

Gene chips have also been classified by function (Hencke, 1998b). Three basic classes have been delineated. The first type is a sequencing chip that serves as a template for analyzing unknown DNA sequences. The second type is categorized as an expression chip. These are designed to determine the degree of expression of a certain genetic sequence by measuring the rate or amount of messenger ribonucleic acid being produced by the target gene. The third type of chip is devoted to comparing the degree of genomic hybridization on the chip produced by different samples of DNA.

7.3 Application of gene chips to basic research

Perhaps the most extensive effort using cDNA microarrays was the first published report of gene expression of an entire genome, that of the yeast, *Saccharomyces cerevisiae.*

DeRisi et al. (1997) studied changes in gene expression during the shift from anaerobic to aerobic metabolism occurring in this organism. Briefly, yeast open reading frames, amplified by polymerase chain reactions, were analyzed with DNA microarrays containing about 6400 distinct DNA sequences printed onto glass slides by a robotic printing device. Exponentially growing cells were harvested at 9 hours and at 2 hour intervals thereafter and mRNA was isolated at each interval. Fluorescently-labeled cDNA was prepared by reverse transcription in the presence of green- and red-labeled deoxyuridine triphosphate (dUTP) and then hybridized to the microarrays. Subsequently, data from the series of seven samples consisting of more than 43,000 expression-ratio measurements were organized into a database for analysis. These investigators found that gene expression in glucose-rich medium, is remarkably stable. Very few changes in gene expression were noted. However, when glucose was depleted from the growth medium, marked changes in expression were noted (e.g. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of approximately 2). The authors noted that about half of these differentially expressed genes had no currently recognized function and were not, as of yet, named. They concluded that DNA microarrays provided a simple and economic way to explore gene expression patterns on a genomic scale, indicating that even simple experiments such as this one, can yield vast amounts of data. In addition, they warned that perhaps the greatest challenge lay not in the use of microarrays, but in the effective organization and analysis of the large amounts of data generated.

Oligonucleotide chips have also been used to evaluate gene-expression patterns in nearly all open reading frames of *S. cerevisiae* during anaerobic and aerobic metabolism (Wodicka et al., 1997). Cellular poly (A)+ mRNA was extracted from yeast cells grown in rich and minimal media, and hybridized to a set of four arrays that contained more than 260,000 specifically chosen oligonucleotides synthesized using light-directed combinatorial chemistry. More than 87% of all yeast mRNAs were detected in yeast cells grown in rich medium. Approximately 19% of all RNAs were present at less than 0.1 copies per cell, 50% between 0.1 and 1 copy per cell, 25% between 1 and 10 copies per cell and 5% at more than 10 copies per cell. The overall distribution of mRNA levels was very similar for cells grown in minimal medium. As in the study using cDNA arrays discussed above, a small number of genes were identified with dramatically different expression levels in rich and minimal media. The authors concluded that the combination of highly parallel detection and monitoring methods (such as oligonucleotide arrays), mutant strains, and use of genetic screens provide a very powerful basis for investigation of yeast biochemistry. They envisioned research findings in yeast utilizing these methods as contributing to a greater understanding of genetic mechanisms in more complex organisms.

Nuwaysir et al. (1999) have listed numerous additional examples of recent research utilizing either cDNA or oligonucleotide microarrays. cDNA microarray investigations include those of *Arabidopsis thaliana* RNA, yeast RNA, yeast genomic DNA, tumorigenic versus non-tumorigenic human tumor cell lines, human T-cells, and human inflammatory disease-related genes (Schena et al, 1995, 1996; DeRisi et al., 1996; Shalon et al., 1996; Lashkari et al., 1997; Heller et al., 1997). Alternatively, oligonucleotide microarrays have been

Tab. 16: Signature correlation of expression ratios as a result of FK506 treatment in various mutant strains (from Marton et al., 1998).

	wild-type +/- FK506	cna +/- FK506	fpr1 +/- FK506	cna fpr1 +/- FK506	cph1 +/- FK506
wild-type +/- FK506	0.93 ± 0.04	-0.01 ± 0.07	-0.23 ± 0.07	0.12 ± 0.07	0.79 ± 0.03

Signature correlation shows the absence of the FK506 signature specifically in the calcineurin (cna) and fpr1 (major FK506 binding protein) deletion mutants. cna represents the mutant with deletions of the catalytic subunits of calcineurin, CNA1 and CNA2. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- FK506 experiments.

applied to mapping of genomic library clones, used to perform *de novo* sequencing by hybridization, and to compare evolutionary sequences of the BRCA1 gene (Sapolsky and Lipshutz, 1996; Pease et al., 1992; Chee et al., 1996; Hacia et al., 1998). Recently, single nucleotide polymorphisms in the human and yeast genomes have also been investigated using oligonucleotide chips (Wang, et al., 1998; Winzeler, et al., 1998).

7.4 Application of gene chips to toxicological research

The melding of genetics and toxicology through the vehicle of DNA microarrays appears to offer numerous opportunities for enhancing toxicological research efforts. Nuwaysir et al. (1999) believe that the microarray approach …"is a tool of unprecedented power for use in toxicology studies". They argue that almost without exception, gene expression is altered during toxic exposure, either directly or indirectly and that toxicologists need to define specific patterns of gene expression elicited by a given toxicant [by using microarrays]. Pennie et al. (2000) concur suggesting numerous dividends to be obtained through the application of genomics to toxicology (e.g. facilitation of more rapid screens for compound toxicity; provision of new research leads; a more detailed appreciation of molecular mechanisms of toxicity). Moreover, they divide toxicogenomics into two classes: mechanistic or investigative research and predictive toxicology. The first is concerned primarily with a mechanistic understanding of the toxic process and the latter with gathering information for predictive purposes. Definition of signature patterns of gene expression following toxicant exposure, as suggested by Nuwaysir above, is considered an example of the use of toxicogenomics for predictive purposes by this classification scheme.

A number of toxicological research studies have already been initiated using DNA microarrays, the majority of which have investigated differences in gene expression following toxicant exposure.

Marton and colleagues (1998) investigated drug target validation using microarrays. They contend that confirmation that a compound inhibits the intended target and identification of undesirable secondary effects are the main challenges in developing new drugs. They evaluated the effect of the drug, FK506 on gene expression in the yeast, *S. cerevisiae*, specifically that of the calcineurin signaling pathway. They showed that mutant strains lacking proteins inhibited by FK506 did not exhibit the "signature" of altered gene expression noted in normal strains (Tab. 16). "Off-target" effects of FK506, that were independent of the drug's binding to immunophilins, were also demonstrated.

Integration of DNA microarrays into the drug development process as suggested by work such as Marton and coworkers, has spawned a number of commercial products most of which look at changes in patterns of gene expression over a large number of genes. One example is ToxExpress™ (Gene Logic, Inc.). This system combines a reference database of expression profiles of rat (tissues and primary cells) and human (primary cells) samples that have shown toxic responses to known compounds with high-throughput systems for gene expression profiling (i.e. microarrays) and software tools. Gene expression profiles for candidate compounds are compared with toxin-associated patterns in the reference database. Pennie et al. (2000)

Tab. 17: ToxChip v1.0: A human cDNA microarray chip designed to detect repsonses to toxic insult (from Nuwaysir et al., 1999).

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PHA responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

This list is intended as a general guide. The gene categories are not unique and some genes are listed in multiple categories.

have referred to this comparison of changes in gene expression with reference profiles as the "pattern recognition" approach to toxicology.

Comparison to a reference set of "toxicant signature" gene patterns is also a major feature of ToxChip v1.0 (Nuwaysir et al., 1999). 2900 human genes compose ToxChip v1.0 including DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress (Tab. 17).

Transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes are also present. 84 "housekeeping" genes that are used to normalize signal are on the chip as well. A hypothetical toxicity protocol using ToxChip v1.0 is as follows. Dose- and time-course parameters would be established for a series of toxicants within a given prototypic class (e.g. polycyclic aromatic hydrocarbons (PAHs). Cells would then be treated with these agents at a fixed toxicity level (as measured by cell survival), RNA would be harvested, and toxicant-induced gene expression changes would be evaluated by hybridization to ToxChip v1.0. When the "toxicant signature" is determined, the genes within this signature are flagged and subsequent experiments with chemicals of unknown action rely on comparison to these patterns established for each prototypic class of chemicals.

Another approach to assessing toxicant exposure through changes in gene expression utilizes smaller microarrays, upon which selected gene sequences have been placed (e.g. ToxBlot arrays) (Pennie et al., 2000). ToxBlot arrays of both human and mouse genes have been constructed which contain, in each case, about 2400 cDNA sequences, representing about 600 genes. Broad gene classes representing cancer, immunology, endocrinology and neurobiology, as well as sequences thought to be useful to investigative toxicology, predevelopment toxicology, and safety assessment are integrated into these arrays (Tab. 18).

A database on gene function, distribution, and allelic variations for each gene represented has also been established. Research investigations of toxicant action on several target organ systems have been performed using ToxBlot arrays. Treatment of cultured cells with natural, synthetic, and phytoestrogens at several doses and time points resulted in a pattern of consistent gene expression changes. Similarly, hepatic toxicity following exposure of HepG2 cells to ethanol, paracetamol, hydrogen peroxide and carbon tetrachloride was characterized by transcript profiling at selected doses for all four compounds. Consistent changes in gene expression were observed here also.

7.5 Development of gene chips as alternatives to animal testing

The relationship of gene chips to alternatives research can perhaps be best described as one of tantalising potential. As indicated by several of the researchers whose work was described above, DNA microarrays can be a very powerful tool, particularly when used to link changes in gene expression with changes in biochemical processes. That this tool can be used effectively in toxicological investigations is evident from work monitoring changes in gene expression following exposure to toxicants. However, how can gene chips contribute to the reduction or replacement of animals in medical ex-

Tab. 18: Broad gene classes included on the ToxBlot Microarrays (from Pennie et al., 2000)

Cancer	Immunology	Endocrinoloy and neurobiology	Investigative toxicology	Predevelopment toxicology	Safety assessment
Apolipoprotein genes	Basic transcription factors	Basic transcription factors	Acetyl CoA pathway	Bcl/Bax family	Bcl/Bax family
Basic transcription factors	Cell adhesion molecules	CYP	Bcl/Bax family	CYP	CYP
Bcl/Bax familiy	Cell surface receptors	Drug metabolism	Drug metabolism	Acetyl CoA pathway	Drug metabolism
CYP genes	Chemokines	Extracellular matrix	Immediate early genes	Ion channels	GST
Caspases	Extracellular matrix	GABA receptors/ transport	GST	GST	Heat shock proteins
cdc/cdk s	Heat shock proteins	Heat shock proteins	CYP	Histones	Liver acute-phase markers
Cyclins	Interleukins	Ion channels	Basic transcription factors	Heat shock proteins	Markers for GI tract physiology
GST	Metalloproteinases	Neurotransmitter-metabolising enzymes	Steroid hormone receptors	Steroid regulated genes	Oxidative-stress markers
Heat shock proteins		Neurotrophic factors/ receptors			Steroid regulated genes
Immediate early genes		Peptide hormones			Thyroid hyperplasia markers
Interleukins		Steroidogenesis/ aromatase			
Matrix metalloproteins		Steroid hormone receptors			
Steroid hormone receptors					

perimentation? The answer(s) lies in the integration of gene chips effectively into prescreening protocols for unknown compounds, in the coupling of gene chips with cell and tissue culture models, and, in a better understanding of the genetic effects of toxicant insult.

Nuwaysir et al. (1999) emphasizes that although the field of toxicology uses many *in vivo* model systems, including rats, mice and rabbits, to evaluate potential toxicity, a variety of *in vitro* techniques already exist to measure toxicity, many of which measure toxicant-induced DNA damage (e.g. Ames test, Syrian hamster embryo cell transformation assay, micronucleus assays, unscheduled DNA synthesis assays). All these methods are predicated upon alterations in gene expression. Hence, gene chips are yet another addition to, if not a replacement for, this collection of *in vitro* assays. Assays that measure DNA damage are already an integral part of most toxicology prescreening programs. The gene chip pattern recognition approach could also most definitely be incorporated into many prescreening batteries for new compound development (e.g. new drug development).

Pennie et al. (2000) suggest that the throughput requirements of pattern recognition approaches using gene chips will…"almost certainly necessitate employing *in vitro* culture systems." Their work (discussed above) did just that, employing both estrogenic and hepatic cells in culture. However, they caution, that although practical advantages to cell cultures, exist, there are major disadvantages. Compound-induced changes in transcription many not necessarily reflect accurately the response of the corresponding organ *in vivo* and availability of cell lines may be limited (particularly if mechanistic endpoints are sought). Also of note is the fact that *in vitro* modeling is unlikely to be an effective replacement for animals when species, strain, sex, or route of administration are important factors in development of the toxic response.

Despite the limitations listed above, much recent research has combined cell cultures with DNA microarrays. Holden and coworkers (1999) have analyzed altered hepatocyte gene expression following exposure to carbon tetrachloride by using human hepatoma cells (HepG2 cell line). A total of 580 genes were analyzed; 40 were found to exhibit modulated expression after exposure to CCl_4 (Tab. 19).

Tab. 19: Selected genes found to be differentially expressed in HepG2 cells following treatment with CCl_4 for 8 hours. Results given as fold induction relative to the DMF control from a single experiment (from Holden et al., 1999).

Gene	Description	Induction (fold)
IL-8	Chemotactic factor/neutrophil activation	+7
IL-1ß	Immunoregulatory cytokine	+3.8
c-fms/csf-1	Tyrosine-protein kinase	-12.8
p55CDC	Cell division control	+4
STAT6	IL-4 signalling	-5.5
STK2	Serine/threonine-protein kinase	+7
RAB-6	Ras-related protein	+8
Apopain/Caspase-3	Executes apoptosis	+4
Bcl-X/Bax	Accelerates apoptosis	+6.6

Rodi et al. (1999) evaluated the effect of phenobarbital in rodent hepatocytes. More than 300 genes were found to exhibit modulated expression after this exposure. Mouse liver cells have been used to evaluate changes in expression of cytochrome P4501A1 and 1A2 genes after beta-naphthoflavone exposure (Bartosiewicz et al., 2000). Upregulation of CYP1A2 was noted at doses lower than those detected using Northern blotting techniques. A human bladder epithelial cell line has been used to assess the effects of arsenite on cell proliferation (Simeonova et al., 2000). Expression of genes associated with cell growth (e.g. c-fos, c-jun, and EGR-1) and cell arrest (e. g. GADD153 and GADD45) was modulated. Furthermore, the National Institutes of Environmental Health Sciences (NIEHS) is currently soliciting grant proposals for development of alternatives to animals for toxicity testing specifically those which combine cell culture and microarray technology (NIEHS, 2000). Proposals are to include means by which "it is possible to assess gene expression in cells in culture and cells *in vivo* and to then devise methods to make the cell cultures mimic the *in vivo* situation at least with regard to gene expression." Hence, the combination of cell culture and DNA microarray technologies is being considered seriously as a method by which replacement alternatives to animal testing can be derived.

Although utilizing microarrays as prescreens of toxicant action can easily be considered a useful reduction alternative, establishing the relationship between changes in gene expression and resulting toxic response may be key to acceptance of assays utilizing microarrays as replacement alternatives. Pennie et al. (2000), however, warn that establishing this linkage can be fraught with difficulties and may require assistance from proteomic technology in order to examine parallels between gene and protein expression modulation after toxic exposure. They suggest that microarray experiments pose a question of "certainty" (i.e. are changes in gene expression 'real', and if so, does altered gene expression translate into protein production?). They also argue that changes in gene expression that do not result in altered protein production are of relevance and should not be dismissed. Indeed, they challenge us to adopt a broader view of biological relevance, one that accepts greater intellectual discomfort, in order to embrace a more holistic perspective of biological change. In essence, events transpiring on a gene chip may not always be reduced/refined to discrete biochemical pathways. They should nevertheless, be considered valuable.

7.6 Current status of gene chip technology

Hencke (1999) summarizes the rapid progress of chip technology as occurring in four areas: sensitivity, microfabrica-

tion, density, and integration. Increased sensitivity is being sought through better labeling technologies such as time-resolved fluorescence, dendrimers, iodine radioisotopes, and chemical locks. A balance will need be struck, however, between the complexity of the testing procedures and the sensitivity of the tests. Microfabrication of chips is already in existence as indicated by the presence of pico- and microliter-sized compartments; nanoscale devices are now also under construction. Increasing the density of the arrays packed onto DNA chips is an objective that is rapidly being reached. Single-chip arrays may now contain 20.000 wells of 100 nl volume; newer technology promises to make available single-molecular layer DNA bound to thin films of gold. An integrated chip, also known as a "lab-on-a-chip" seems to be where much work in chip technology remains to be done. These chips integrate sample preparation and analysis, reducing the "complete system" as designated by Botwell above, to two or fewer components. Improved microfluidic capabilities to capitalize on the multiple uses of capillary electrophoresis as well as on-chip chemiluminescence, and electrochemiluminescence assays are being honed for integration into chip manufacture.

Many of the research possibilities for gene chips have yet to be explored and novel ideas for their use continue to spring forth. However, major hurdles currently face manufacturers and users of the chips, the primary example of which is quality.

Government regulators and users will have to be guaranteed that the chips are of the highest quality possible. Because the effective manufacture of DNA microarrays is still undergoing development/refinement, quality control issues continue to require re-definition. Several large companies have launched a group called the Genetic Analysis Technology Consortium (GATC) to evaluate these issues. This group intends to standardize array-based genetic analysis, "paving the way for more-affordable and productive development of diagnostic and therapeutic products" (Hencke, 1999).

In addition to issues regarding gene chip quality, concerns surrounding the economics of gene chip use need to be addressed. A new technology may provide unique solutions to current problems, but if it's not affordable, no one will use it. Marketing gene chips as a less expensive option to current diagnostic procedures is probably the only way in which they will be accepted in the current marketplace. Unfortunately, making gene chip manufacture less costly will require the development of cost-effective techniques for both information management and sample preparation components considering the large quantities of data requiring analysis produced from these experiments. This will not be an easy task.

7.7 Summary, conclusions, and future work

In summary, gene chips (DNA microarrays) represent a technology that has already opened many doors in basic genomic research. Moreover, their value to both investigative and discovery toxicology is becoming much more apparent as more toxicology experiments are conducted using them. Use of microarrays as reduction or replacement alternatives to animal testing also holds great promise, particularly when they are used as components of prescreening batteries, and when coupled to cell culture techniques.

In the next five years research efforts directed toward use of "toxicant signatures" and pattern recognition approaches to develop reproducible screening techniques for structurally similar groups of chemicals will probably be the most fruitful. This work may also be extended to include groups of genes involved in particular target organ toxicity. Either option will likely produce a method for use as a reduction alternative that can be integrated into prescreening batteries. General changes in gene expression will be monitored and reproducible patterns will be sought. *In vitro* assays in which a mechanistic relationship between changes in gene expression and toxicant insult is well defined should be validated in the next ten years. Regulatory acceptance of this type of assay should occur during that time as well. In the next twenty years, many of the genes affected by toxicant action should be identified supporting efforts focused on understanding mechanisms of action. When accurate parallels can be drawn between changes in gene expression and exposure to toxic chemicals, gene chips will, without a doubt, function admirably as replacements for animals in toxicity testing.

References

Bartosiewicz, M., Trounstine, M., Barker, D. et al. (2000). Development of a toxicological gene array and quantitative assessment of this technology. *Arch. Biochem. Biophys. 376(1)*, 66-73.

Bowtell, D. (1999). Options available: From start to finish – for obtaining expression data by microarray. *Nature Genetics (Supplement) 21*, 25-32.

Brown, P. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics (Supplement) 21*, 33-37.

Chee, M., Yang, R., Hubbell, E. et al. (1996). Accessing genetic information with high-density DNA arrays. *Science 274*, 610-614.

DeRisi, J., Penland, L., Brown, P. et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics 14*, 457-460.

DeRisi, J., Iyer, V. and Brown, P. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science 278*, 680-686.

Ermolaeva, O., Rastogi, M., Pruitt, E. et al. (1998). Data management and analysis for gene expression arrays. *Nature Genetics 20*, 19-23.

Gene Logic, Inc. (2000). http://www.genelogic.com/products/white-tox.pdf

Hacia, J., Makalowski, W., Edgemon, K. et al. (1998). Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nature Genetics 18(2)*, 155-158.

Heller, R., Schena, M., Chai. A. et al. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA 94*, 2150-2155.

Hencke, C. (1998a). DNA-chip technologies. Part 1: Research fundamentals and industry catalysts. *IVD Technology.*

Hencke, C. (1998b). DNA-chip technologies. Part 2: State-of-the art and competing technologies. *IVD Technology.*

Hencke, C. (1999). DNA-chip technologies. Part 3: What does the future hold? *IVD Technology.*

Holden, P., James, N., Brooks, A. et al. (1999). Analysis of altered hepatocyte gene expression induced by carbon tetrachloride (CCl4) using microarray technology. *Hum. Exp. Toxicol. 18(8)*, 522.

Lashkari, D., DeRisi, J., McCusker, J. et al. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA 94*, 13057-13062.

Lockhart, D., Dong, H., Byrne, M. et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology 14*, 1675-1680.

Marton, M., DeRisi, J., Bennett, H. et al. (1998). Drug target validation and identification of secondary drug target effects using DANN microarrays. *Nature Medicine 4(11)*, 1293-1301.

NIEHS (National Institute for Environmental Health Sciences) (2000). Solicitation of the Public Health Service for Small Business Innovation Research Contract Proposals. PHS 2001-1. National Institute for Environmental Health Sciences (NIEHS). Section X. Research Topics: 079 Development of Alternatives to Animals for Toxicity Testing. p. 37.

Nuwaysir, E. F., Bittner, M., Trent, J. et al. (1999). Microarrays and toxicology: The advent of toxicogenomics. *Molecular Carcinogenesis 24*, 153-159.

Pease, A., Solas, D. and Fodor, S. (1992). Parallel synthesis of spatially addressable oliognucleotide probe matrices. Abstract. *American Chemical Society. 203*, 34.

Pennie, W. D., Tugwood, J., Oliver, G. and Kimber, I. (2000). The principles and practice of toxicogenomics: applications and opportunities. *Toxicological Sciences 54*, 277-283.

Rodi, C., Bunch, R., Curtiss, S. et al. (1999). Revolution through genomics in investigative and discovery toxicology. *Toxicol. Pathol. 27*, 107-110.

Sapolsky, R. and Lipshutz, R.(1996). Mapping genomic library clones using oligonucleotide arrays. *Genomics 33*, 445-456.

Schena, M., Shalon, D., Davis, R. and Brown, P. (1995). Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science 270*, 467-470.

Schena, M., Shalon, D., Heller, R. et al. (1996). Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA 93*, 10614-10619.

Shalon, D., Smith, S. and Brown, P. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research 6*, 639-645.

Simeonova, P., Wang, S., Toriuma, W. et al. (2000). Arsenic mediates cell proliferation and gene expression in the bladder epithelium: association with activating protein-1 transactivation. *Cancer Res. 60(13)*, 3445-3453.

Wang, D., Fan, J., Siao, C. et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science 280*, 1077-1082.

Winzeler, E., Richards, D., Conway, A. et al. (1998). Direct allelic variation scanning of the yeast genome. *Science 281*, 1194-1197.

Wodicka, L., Dong, H., Mittmann, M. et al. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnology 15*, 1359-1367.

8 Validation of alternative methodology

8.1 Background

In an early article about cell culture alternatives to animal testing, Goldberg (1986) wrote that, "The need is stressed for adequate standardization and validation of CCS (cell culture systems), and meticulous conduct of tests using such systems." Mehlman et al. (1989) expanded these sentiments in a later report on methods to reduce, refine and replace animal testing in industrial laboratories. They stated that, "There is a pressing need for the systematic and scientifically sound validation of non-animal alternatives techniques to reduce the use of animals in toxicology testing while satisfying requirements for the protection of public safety". Both statements reflect the fact that validation of test methods has been recognized early on as essential to the development and subsequent acceptance of any alternative method by many researchers in this area.

Gad (1990) subsequently wrote that a scientific consensus had been formed on the requirements and process for validation (of alternative methods). Although this consensus might be considered debatable at times, serious attempts at standardizing the requirements of validation and the process by which it can be achieved have been performed (Balls et al., 1990; Balls et al., 1995a; Balls et al., 1995b). Moreover, a research facility specifically designed to investigate the validation of alternative methods, the European Center for the Validation of Alternative Methods (ECVAM), has been established. Figure 14 contains a diagram of a proposed validation process.

Furthermore, Goldberg et al. (1993) suggested that, "A framework capable of fostering the validation of new methods is essential for the effective transfer of new technology from the research laboratory to practical use". An example of part of this framework in place in the United States is the recent establishment of the Interagency Coordinating Committee on the Validation of Alternative

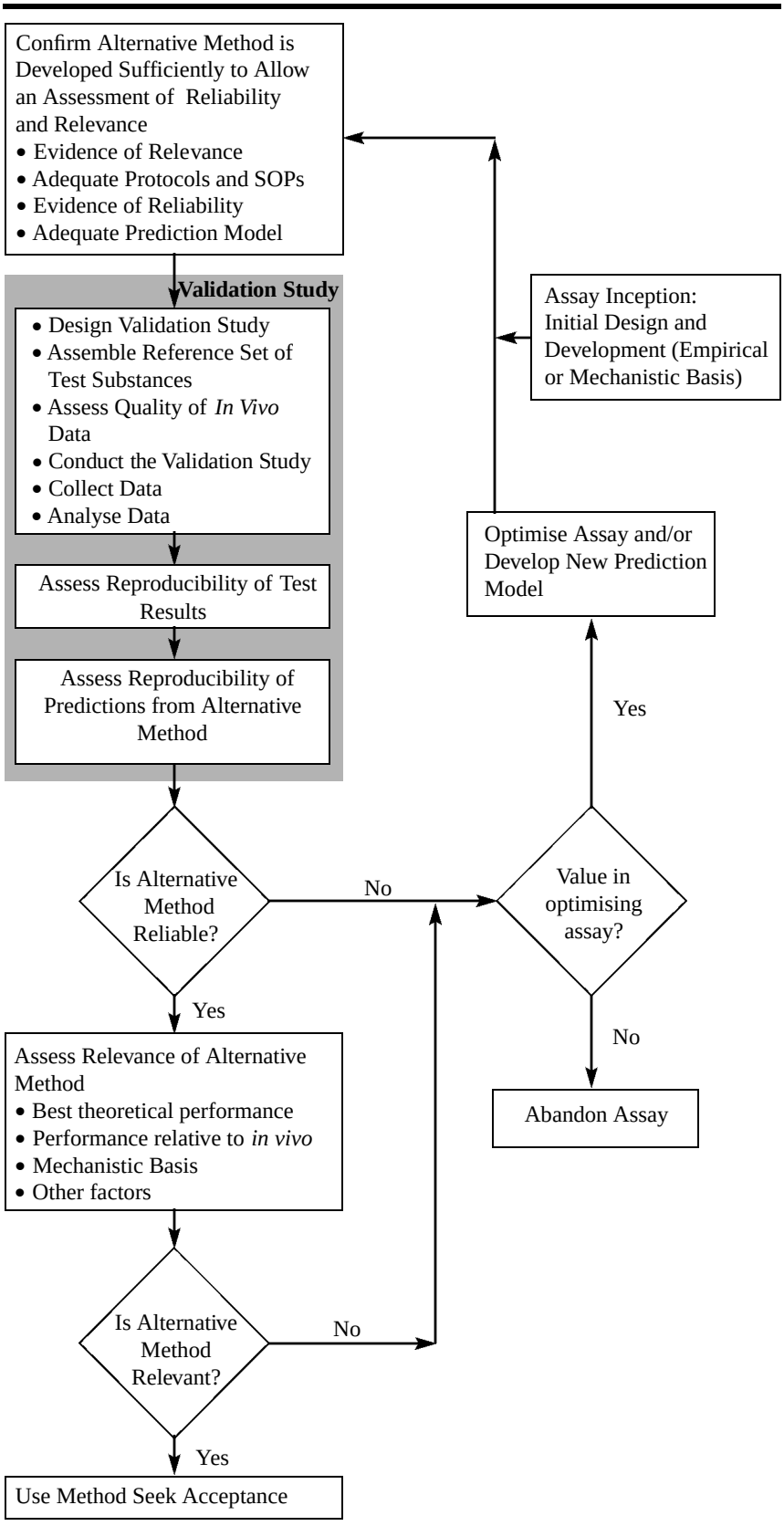

Fig. 14: Validation process (from Bruner et al., 1996).

Methods (ICCVAM) working in concert with the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). This type of linkage is also a strong indication that regulatory bodies have recognized the validation process as integral to effective development of a viable alternative method.

8.2 What is validation and why has it failed?

Definitions of validation vary. However, the following excerpts represent fundamental descriptions of validation as they apply to alternatives to animal testing. Balls et al. (1990) state that, "Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose". Moreover, a later perspective indicates that, irrespective of purpose, or the type of validation study, the key factor is that it is of a scientific, rather than a political, nature. (Balls et al., 1995a). This latter view emphasizes an important distinction that is necessary, although often overlooked, for the maintenance of objectivity in a regulatory arena that often exhibits strongly political overtones.

Despite the realization that validation is an important part of the development and ultimate regulatory acceptance of an alternative method, numerous problems have arisen with the appropriate conduct of validation studies. Balls and coworkers (1995a) compiled the following list of notable general deficits:

- The goals of the validation study were not sufficiently well defined, and there were differences of opinion as to what the objectives should be, and how best to achieve them.
- Many of the studies were poorly designed and planned.
- The responsibilities of the participants have not been clearly defined and/or fulfilled.
- The validation study has been inadequately managed due to, for example, funding and/or communication problems.
- Standard protocols defining all aspects of the test procedure have been lacking, or have not been strictly adhered to by some of the participating laboratories.

- Relevant test chemicals, for which high quality and unambiguous *in vivo* (animal and/or human) toxicity data were available, could not be identified or obtained.
- Participating laboratories have failed to meet the deadlines set for completion of the testing, submission of data, and other critical stages of the study.
- Participants in the validation study have been unable to agree on how to evaluate the data or on the interpretation of the test results.
- Laboratories have committed themselves to participating in a validation study, but have subsequently dropped out once the study was under way, for financial or logistical reasons.

Furthermore, reasons for the failure of several specific validation studies of alternatives to eye irritation testing have recently been investigated (Balls et al., 1999). The following conclusions were drawn:

- The *in vivo* test to which most of *in vitro* alternatives were compared provided highly variable results due to a subjective scoring system
- Many non-animal protocols were inadequate
- The choice of test substances was not rigorously defined
- The choice of statistical methods for data analysis was not as robust as it could have been

Although several of the conclusions drawn from this investigation were similar to the more general ones listed above, it should be noted that most *in vivo* models of toxicity testing do not exhibit as much variability as that observed with the Draize eye irritation test. Therefore, this particular validation problem may apply only to attempts to validate eye irritation alternatives.

As judged from these investigations, the *in vitro* assay undergoing validation may fail due to causes that have less to do with the actual value of the assay than with the conduct of the validation study.

In an additional critique of the validation process, Koeter (1995) wrote that validation is a highly charged concept understood as necessary by many, but unable to produce more widespread acceptance of alternative methods. He cited two reasons for this failure: "First, the results of the validation studies may have been unsatisfactory, which could mean that either the method subjected to validation failed to show the desired relevance and reliability, or the validation study as such yielded inconclusive results. Secondly, despite clear-cut (supporting) results from the validation exercise, toxicologists/regulators appear reluctant actually to use the data provided for hazard and risk assessment procedures because of a lack of confidence with the (types of) endpoints of the new test." He considered the latter reason to be the "major hurdle" to the acceptance of alternative methods. However, his first reason, in synchrony with the views of other researchers, emphasizes the need for adequate conduct of validation studies.

8.3 Improving the validation process

General remedies have been advanced for improving the validation process (Balls et al., 1995a). Suggestions include:

- A clear and unequivocal statement of what the validation study is designed to accomplish
- A well-defined plan for the study
- A sufficiently large set of test substances covering the relevant chemical classes, the different categories of cosmetic ingredients, and the range of toxic endpoints to be evaluated
- *In vivo* data of high quality on all the test substances to be used
- Evidence that the methods to be evaluated are scientifically-sound, relevant, reproducible, and that they have the potential to replace the animal test in question
- An optimized protocol for each test, with any necessary standard operating procedures
- A clear description of how each alternative method can be used to predict an *in vivo* endpoint
- Agreed statistical procedures for testing whether the methods can predict the *in vivo* endpoints defined by their developers
- Agreed criteria to be met in order to show that an alternative method could successfully and safely replace an animal test
- Results which meet these criteria

Moreover, specific suggestions for enhancing the validation of eye irritation alternatives have been made (Balls et al., 1999). These included the following:

- A new approach to the validation of *in vitro* tests for eye irritancy, based on the use of reference standards
- The use of stepwise testing strategies which reduce and refine the use of animals in eye irritation testing
- The use of multivariate and other statistical techniques for the further analysis of data generated in previous validation studies
- A program of research aimed at understanding the underlying mechanisms of eye irritation

Hence, ways to improve the validation process in general as well as specific suggestions aimed at a particular type of validation study, that of the eye irritation alternative, have been put forth.

In addition, perhaps one of the most important concepts suggested as a means of improving validation efforts has been that of "prevalidation". Prevalidation was proposed to: (1) standardize test procedures into a formal protocol; (2) establish multiple laboratories competent in performing the test method in line with Good Laboratory Practice (GLP) principles; (3) develop the test's prediction model (the algorithm which converts *in vitro* data to a predicted *in vivo* hazard) (Fig. 15); and (4) supply independent evaluation of the performance of the test protocol and the prediction model (Curren and Southee, 2000).

Curren et al. (1995) defined prevalidation as consisting of three phases. Phase I was defined as Protocol Development in which a GLP-compliant protocol and Standard Operating Procedures (SOPs) would be created, and determination of intralaboratory reproducibility would be accomplished. If successful, Phase I activities would then be followed by Phase II, Protocol Transfer. Protocol Transfer would consist of transfer of the method to a second laboratory using the defined protocol and SOPs, determination of interlaboratory reproducibility, further refinement of the protocol if needed, submission of the optimized protocol to a documentation source, e.g. INVITTOX, and evaluation of suitability

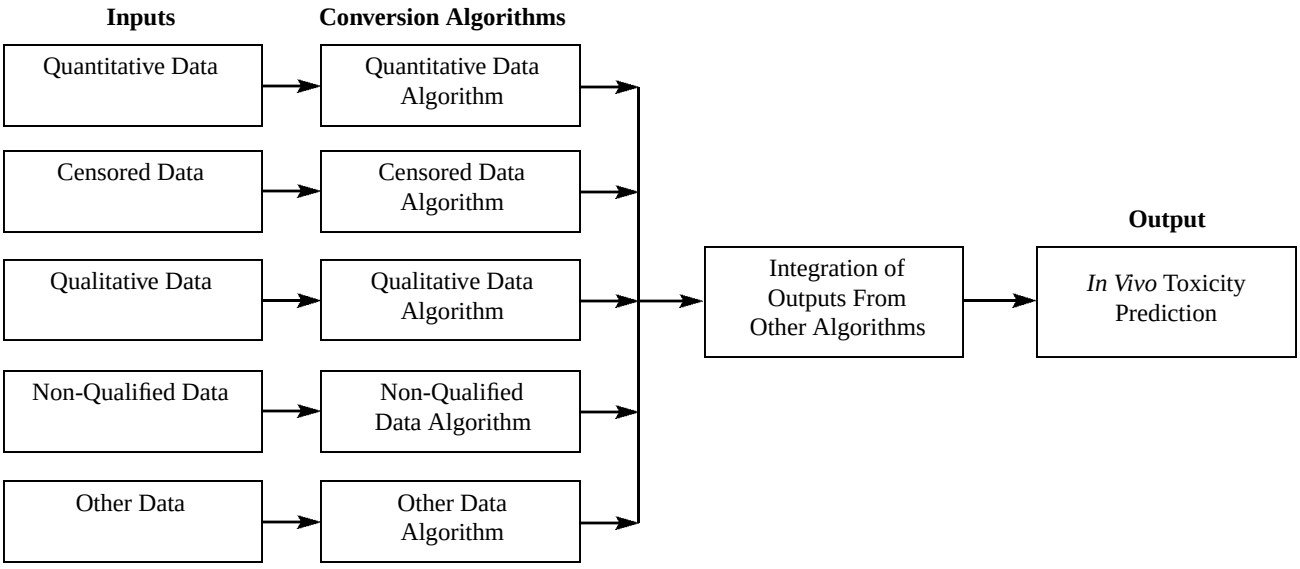

Fig. 15: Converting alternative method data into an *in vivo* toxicity prediction (from Bruner et al., 1996).

for progression to Phase III. Phase III, Protocol Performance, would require transfer of the method to a third laboratory, testing of coded materials in at least three laboratories, preparation of a report on the performance of the method, and confirmation or redefinition of the prediction model.

The concept of prevalidation has been evaluated by applying it "in practice" to several *in vitro* assays including the embryonic stem cell (EST) assay, the bovine corneal opacity and permeability (BCOP) assay, the fluorescein leakage test, the EpiDerm™ Skin Corrosivity test, and the EpiDerm™ phototoxicity test. Recent applications have also included the transcutaneous electrical resistance (TER) and Episkin™ assays as well as the 3T3 NRU assay. Application of prevalidation procedures to evaluation of these assays was successful; several of them have been approved as scientifically valid by the ECVAM Scientific Advisory Committee (ESAC) (Balls and Corcelle, 1998a; Balls and Corcelle, 1998b; Balls and Corcelle, 1998c). Curren and Southee advise, however, that effective communication is critical to successful prevalidation efforts as is the "spirit of cooperation". They note that one of the most persistent problems in these prevalidation efforts, as with full-blown validation attempts, was obtaining an adequate number of test chemicals with acceptable *in vivo* data with which to conduct a prevalidation exercise.

A specific area of both prevalidation and validation studies which has also come under close scrutiny is that of the biostatistical techniques used to draw *in vitro/in vivo* correlations. These techniques represent one of the most heavily weighted factors used in determining whether sufficient correlation exists between *in vitro* and *in vivo* methods to justify saying that the *in vitro* method is predictive of *in vivo* toxicity. In an attempt to provide some basis for standardization of these techniques, Bruner et al. (2000) address this issue when describing objective assessment of test method performance in validation studies. They state that, "In order to make further progress in developing validation methodology, two topics that need further exploration are how to adequately measure the predictive capacity of toxicity tests evaluated in validation studies, and what values obtained from these measurements of predictive capacity are indicative of acceptable test method performance." They suggest that this issue deserves special consideration in view of the numerous data types derived from the toxicity test methods under evaluation and the plethora of statistical techniques which have been used to assess performance of the test method. They analyzed four simulated validation study data sets using three different approaches: calculation of correlation coefficients, determination of the 95% prediction interval (PI) and calculation of sensitivity, specificity and accuracy as defined by Cooper et al. (1979). They found that calculation of the 95% PI provides the most information regarding performance of the test method under study. They also found that dividing data sets into positive and negative toxicity classifications followed by the calculation of sensitivity, specificity and accuracy, leads to significant information loss.

8.4 Harmonization of validation efforts

As indicated above, improvement and standardization of validation efforts has received vigorous attention in the last five years. This attention was felt to be justified in view of the fact that alternatives to eye irritation testing in animals, which have been in existence the longest

and about which the majority of data has been generated, could not be validated to the extent necessary to assure regulatory acceptance in Europe or the United States. Although much work remains to be accomplished regarding the validation of eye irritation alternatives, improvements made to the validation process resulting from this inquiry, particularly those involving a "prevalidation" of the alternative test method, have been applied successfully to areas of research other than those of eye irritation alternatives (for example, skin corrosivity).

Fentem (2000) in noting the progress made toward effective validation of alternative methods observed that a retrospective approach using data already generated by industrial laboratories or in the scientific literature had been used by ICCVAM to validate the local lymph node assay. Subsequent validation of the Corrositex® assay by this same body, however, relied upon data from formal validation studies. She concluded from this and other validation efforts by ECVAM that validation was now actually "in practice" as a reality as opposed to a theory. She suggested that the procedure for conducting formal validation studies adopted by ECVAM could form the starting point for discussions on international acceptance of validation procedures and their outcomes and that perhaps the OECD (Organization for Economic Cooperation and Development) could oversee this process.

ECVAM's validation procedure (Balls et al., 1995a) consists of the following main points:

- Test development
- Prevalidation (involving three phases as discussed above)
- Validation (involving a formal interlaboratory study)
- Independent assessment
- Progression toward regulatory acceptance

ICCVAM has established its own set of validation steps (NIEHS, 1997) as follows:

- Identify needs for new and/or improved test methods
- Understand toxic mechanisms
- Incorporate new science and technology into test methods
- Optimize transferable protocol
- Determine reliability and relevance
- Independent peer review evaluation of validation status
- Determine acceptability for regulatory risk assessment

As can be seen, much overlap exists between the two methods of validation. The first three steps of ICCVAM's approach are probably encompassed by the first step in ECVAM's strategy. Optimization of a transferable protocol (ICCVAM) is the objective of the prevalidation procedure suggested by ECVAM. Determination of reliability and relevance (ICCVAM) is the outcome of a successful validation study that includes interlaboratory comparisons (ECVAM). Finally, independent assessments by experts are a part of both schemes as is movement toward regulatory acceptance upon successful completion of the previous steps in either strategy. Hence, major differences between the two strategies are not apparent. However, much work probably remains to be done with regard to achieving a consensus about issues of a more technical nature, such as biostatistical methods of evaluation as discussed above. Therefore, efforts directed toward harmonization of validation methods are warranted and suggestions by those such as Fentem above, indicating that ECVAM's approach could perhaps serve as a starting point for discussion, are pertinent.

8.5 Why validate?

Brown (1992) noted that, "It appears that the past 10 years have produced relatively few advances in this use of alternative tests. I will suggest that this is because too much effort has been concentrated on so-called 'validation' studies." It would appear from the discussion above and the comments below, that some of this effort has finally borne fruit, particularly in view of the effective validation of several alternative methodologies through use of improved validation methods.

Fentem (2000) states that, "Principally, validation studies are conducted to provide objective information on new tests, to show that they are robust and transferable between laboratories, and that the data generated can be relied upon for decision-making purposes (for example, for the identification and labeling of a potential skin corrosive, severe eye irritant, teratogen, etc.)."

Others have confirmed the value of validation efforts as follows:

- Curren et al. (1998), in their report of the 13th meeting of the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC), write that "*Scientific principles demand that before newly developed alternative methods for safety testing are fully embraced by the industrial or regulatory community, they reliably and reproducibly predict the designated toxic end point. The process used to determine reliability and reproducibility is termed validation, and it generally culminates with a highly controlled, blinded study using multiple chemicals and laboratories. It is imperative that the validation study is designed to confirm the previously established reproducibility and predictive power of the assay.*"
- Tennant (1998), in his discussion of evaluation and validation issues in the development of transgenic mouse carcinogenicity bioassays, states that, "*Although results in transgenic models may not be completely concordant with long-term bioassays, the data can be used effectively in chemical and drug safety assessments. Further, it is proposed that validation of the models is readily achievable via ongoing studies.*"
- Artiges (1999) in his description of the role of the European Pharmacopoeia in developing alternative methods writes that, "*All these changes [deletion or replacement of animal tests, reduction in the number of animals used and possibilities of refining tests in ways that cause less suffering] require extensive validation work and collaborative trials in a number of control laboratories to ensure that the quality standards are maintained with the alternative methods.*"

Hence, validation seems to have "come of age" after a lengthy process of definition and refinement, despite negative comments to the contrary in the early stages of its development.

8.6 Mechanistic understandings: what are we validating?

One of the primary challenges of developing alternative models of toxicity testing has been the question of relevance of endpoint measured *in vitro* to the toxic response noted in the whole animal. In terms of validation, many parties question the wisdom of validating an endpoint that may or may not have much to do with the basic mechanism of action of a toxic chemical. However, others argue that if a good correlation is achieved between the endpoint measured in an *in vitro* assay and an *in vivo* endpoint, then the *in vitro* assay is indeed predictive of the *in vivo* response. They also point out that determining the mechanism of action of a toxicant is, for most toxicants, a very lengthy procedure.

As noted in an earlier portion of this position paper, a good example of this quandary regarding relevance is recent work with gene chips. Pennie et al. (2000) have argued that changes in gene expression that do not result in altered protein production are of relevance and should not be dismissed. They suggest that we adopt a broader view of biological relevance and realize that events transpiring on a gene chip may not always be reduced/refined to discrete biochemical pathways. They indicate that these events should nevertheless, be considered valuable. This perspective is undoubtedly true for other areas of alternatives to animal testing and should definitely be incorporated it into our thinking about validation efforts.

8.7 Summary, conclusions, and future work

Experiments designed to evaluate how closely results from an *in vitro* assay mirror those obtained from an *in vivo* assay have been performed practically since the beginning of the alternatives to animal testing movement (and much earlier if we look at early experiments performed "*in vitro*" in many disciplines). However, only recently has this type of exercise been subjected to attempts at formalization and standardization. These relatively recent efforts have been prompted by the need to evaluate objectively those *in vitro* assays that could be used to reduce, refine, or replace animals in medical experimentation in order to promote regulatory acceptance of them.

Validation of alternative methods has just emerged from a rather chaotic phase in which the principles behind appropriate conduct of a validation study were defined, mainly through trial and error. Much refinement has come out of this "exploratory" phase, including recognition that validation studies should be build upon a solid platform, consisting of components such as good reference standards, reliable protocol transfer between laboratories, and appropriate application of biostatistical techniques. Efforts are now underway to apply these lessons learned to future validation studies and to harmonize validation techniques among countries in order to maximize the possibility that the data generated can be used worldwide.

So, despite the fact that some *in vitro* assays may not have been validated in the past due to faults inherent in the validation study and not in the assay, future attempts to validate *in vitro* assays will benefit from a better understanding of the requirements of validation and standardization of the conduct of validation studies. Furthermore, validation of endpoints that are not directly correlated with a defined mechanism of action is slowly becoming more acceptable. This is due, in part, to our increasing realization of the fact that our knowledge about mechanisms of action at any given time may be limited. It is hoped that this perspective will be taken into consideration by regulatory bodies charged with the task of evaluating and ultimately, accepting alternatives to animal testing as viable tests of toxicity. If so, Koeter's second reason for the lack of regulatory acceptance of alternatives mentioned above (i.e. "toxicologists/regulators appear reluctant actually to use the data provided for hazard and risk assessment procedures because of a lack of confidence with the (types of) endpoints of the new test.") may be set aside as no longer viable.

Validation, while not an actual alternative to animal testing, certainly can be said to be the "gateway" through which each proposed alternative must pass in order to find acceptance as a viable test of toxicity. Therefore, the methods by which we conduct validation studies are as important, if not more so, than those by which we conduct an alternative assay. Our previous attempts at validation have been well meant, but expensive, and, in some cases, not so well planned or executed. We have, however, through reassessment, honed the technique and applied it successfully to produce validated alternatives to animal testing. All that truly remains to be done is to see that the technique is used consistently to evaluate alternative methods, thereby contributing to a much needed "atmosphere of trust" among scientists, the industrial sector, and regulators.

References

Artiges, A. (1999). Alternatives to animals in the development and control of biological products for human and veterinary use. The role of the European Pharmacopoeia. *Dev. Biol. Stand. 101*, 29-35.

Balls, M., Blaauboer, B., Brusick, D. et al. (1990). Report and recommendations of the CAAT/ERGATT workshop on the validation of toxicity test procedures. *ATLA 18*, 313-337.

Balls, M., Blaauboer, B., Fentem, J. et al. (1995a). Practical aspects of the validation of toxicity test procedures. Report and recommendations of ECVAM workshop 5. *ATLA 23*, 129-147.

Balls, M., De Klerck, W., Baker, F. et al. (1995b). Development and validation of non-animal tests and testing strategies: the identification of a coordinated response to the challenge and the opportunity presented by the sixth amendment to the cosmetics directive. Report and recommendations of ECVAM workshop 7. *ATLA 23*, 398-409.

Balls, M. and Corcelle, G. (1998a). Statement on the scientific validity of the rat skin electrical resistance (TER) test (an in vitro test for skin corrosivity). *ATLA 26*, 275-277.

Balls, M. and Corcelle, G. (1998b). Statement on the scientific validity of EPISKIN™ test (an in vitro test for skin corrosivity). *ATLA 26*, 278-280.

Balls, M. and Corcelle, G. (1998c). Statement on the scientific validity of the 3T3 NRU PT test (an in vitro test for phototoxic potential). *ATLA 26*, 281-283.

Balls, M., Berg, N., Bruner, L. et al. (1999). Eye irritation testing: The way forward. The report and recommendations of ECVAM workshop 34. *ATLA 27*, 53-77.

Brown, N. (1992). Hazard identification in developing toxicology – the role of nonwhole mammal tests. Reprod. *Toxicol. 6(2)*, 178.

Bruner, L., Carr, G. and Curren, R. (2000). Is it good enough? Objective assessment of toxicity test method performance in validation studies. In M. Balls, A.-M. van Zeller and M. E. Halder (eds.), *Progress in the Reduction, Refinement, and Replacement of Animal Experimentation*, (375-384). Amsterdam, London, New York, Tokyo: Elsevier.

Cooper, J., Saracci, R. and Cole, P. (1979). Describing the validity of carcinogen screening tests. *Br. J. Cancer 39*, 87-89.

Curren, R., Southee, J., Spielmann, H. et al. (1995). The role of prevalidation in the development, validation and acceptance of alternative methods. *ATLA 23*, 211-217.

Curren, R., Bruner, L., Goldberg, A. and Walum, E. (1998). 13th Meeting of the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC): validation and acute toxicity testing. *Environ. Health Perspect. 106(2)*, 419-425.

Curren, R. and Southee, J. (2000). Prevalidation: successes justify the concept. In M. Balls, A.-M. van Zeller and M. E. Halder (eds.), *Progress in the Reduction, Refinement, and Replacement of Animal Experimentation*, (359-363). Amsterdam, London, New York, Tokyo: Elsevier.

Fentem, J. (2000). Validation in practice – from theory to reality. In M. Balls, A.-M. van Zeller and M. E. Halder (eds.), *Progress in the Reduction, Refinement, and Replacement of Animal Experimentation*, (365-373). Amsterdam, London, New York, Tokyo: Elsevier.

Gad, S. (1990). Recent developments in replacing, reducing, and refining animal use in toxicologic research and testing. *Fundam. Appl. Toxicol. 15(1)*, 8-16.

Goldberg, A., Frazier, J., Brusick, D. et al. (1993). Framework for the validation and implementation of in vitro toxicity tests: Report of the validation and technology transfer committee of the Johns Hopkins Center for Alternatives to Animal Testing. *J. Am. Coll. Toxicol. 12(1)*, 23-30.

Goldberg, L. (1986). Charting a course for cell culture alternatives to animal testing. *Fundam. Appl. Toxicol. 6(4)*, 607-617.

Koeter, H. (1995). Validation: A highly charged concept. *Toxicol. In Vitro 9(6)*, 851-856.

Mehlman, M., Pfitzer, E. and Scala, R. (1989). A report on methods to reduce, refine and replace animal testing in industrial toxicology laboratories. *Cell Biol. Toxicol. 5(3)*, 349-358.

NIEHS (National Institutes of Environmental Health Science) (1997). Validation and regulatory acceptance of toxicological test methods: A report of the ad hoc interagency coordinating committee on the validation of alternative methods. NIH publication no. 97-3981. Research Triangle Park, NC, USA.

Pennie, W. D., Tugwood, J., Oliver, G. and Kimber, I. (2000). The principles and practice of toxicogenomics: applications and opportunities. *Toxicological Sciences 54*, 277-283.

Tennant, R. (1998). Evaluation and validation issues in the development of transgenic mouse carcinogenicity bioassays. *Environ. Health Perspect. 106(2)*, 473-476.